8-1-2018

CPVIB-1, a GAGA Regulator of TOR Signaling Pathways in the Chestnut Blight Pathogen Cryphonectria Parasitica

Di Ren

CPVIB-1, a GAGA regulator of TOR signaling pathways in the chestnut blight pathogen *Cryphonectria parasitica*

By

Di Ren

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Biological Sciences
in the Department of Biological Sciences

Mississippi State, Mississippi

August 2018

CPVIB-1, a GAGA regulator of TOR signaling pathways in the chestnut blight pathogen

Cryphonectria parasitica

By

Di Ren

Approved:

_____

Angus L. Dawe
(Major Professor)

_____

Donna M. Gordon
(Committee Member)

_____

Matthew W. Brown
(Committee Member)

_____

Andy D. Perkins
(Committee Member)

_____

Mark E. Welch
(Graduate Coordinator)

_____

Rick Travis
Dean
College of Arts & Sciences

Name: Di Ren

Date of Degree: August 10, 2018

Institution: Mississippi State University

Major Field: Biological Sciences

Major Professor: Angus L. Dawe

Title of Study: CPVIB-1, a GAGA regulator of TOR signaling pathways in the chestnut blight pathogen *Cryphonectria parasitica*

Pages in Study 206

Candidate for Degree of Doctor of Philosophy

Cryphonectria parasitica is the causal agent of chestnut blight, which devastated the American Chestnut tree population in the early 20th century. The discovery of hypoviruses that reduce the severity of the chestnut blight infection offers the potential for biological control. However, the spread of the hypoviruses is hampered by a diverse genetically controlled nonself-recognition system, vegetative incompatibility (*vic*). CPVIB-1 was identified as a transcription regulator playing an important role in the programmed cell death response to this stimulus. In this study, we have found that CPVIB-1 is ubiquitin-decorated which might lead to its degradation in the proteasome pathway. RNA-Seq and ChIP-Seq were used to further explore the downstream targets of CPVIB-1 that mediate the various metabolic changes that lead to the altered phenotype of the Δ*cpvib-1* mutant. Due to inaccuracies in the prior annotation, we performed a genome re-annotation to improve the accuracy using a MAKER2-two-pass pipeline. To validate the improvement a second pipeline, PEPA, was developed to compare quality metrics between the old and new annotations. Approximately 1/3 of the original annotations from 2009 were found to be inaccurate. Experimental confirmation by testing 27 predicted

genes using a diagnostic PCR protocol to differentiate between prior and new transcript structures showed that over 80 % of tested genome locations supported for the new annotation. Using rapamycin treatment to mimic stimulation of the *vic* response and applying the RNA-seq and ChIP-seq data to this new information, we found that CPVIB-1 is related to TOR signaling pathways, promoting autophagy and the proteasome pathway, but repressing carbon metabolism, protein and lipid biosynthesis. In depth analysis of CPVIB-1-bound DNA targets showed that this protein is a member of the GAGA regulator family, a group of multifaceted transcription factors with diverse roles in gene activation and repression, maintenance of mitosis, and cell development. Following treatment with rapamycin the recognition sequence bound by CPBVIB-1 was altered leading to the regulation of different suite of genes with diverse metabolic functions. Ultimately, we have developed a revised model of TOR signaling pathway where TORC1 and TORC2 signaling pathways are connected by the action of CPVIB1.

DEDICATION

I would like to dedicate this dissertation to my father Zhihao Ren, my mother Xiulan Zhu and my sister's family as well as all the friends who love and support me. They were always there to cheer me up and stood by me through the good and bad times in last six years while I was pursuing my doctorate degree in biology fields.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my major advisor, Dr. Angus L. Dawe, for his understanding, patience, enthusiasm, and encouragement and for training me to be an independent researcher. I am greatly thankful to Dr. Donna Gordon, Dr. Matthew Brown and Dr. Andy D. Perkins for your assistance and suggestions throughout my project.

I would like to express my sincere appreciation to Dr. Mingzhou Song in Department of Computer Science, New Mexico State University for introducing me to enter the bioinformatics field from the elementary courses and projects. Dr. Brad Shuster, Dr. Immo Hansen, and Dr. Jiannong Xu from New Mexico State University, I am grateful for your help and suggestions at the beginning of the project. To Dr. Peterson, Mark A. Arick, II (*Tony*), Adam Thrash, and Dr. Chuan-Yu Hsu from IGBB, I am thankful for your support to my project in bioinformatics and genomics fields.

For all my lab mates, I am thankful to have your help, support, and fellowship. Especially for Gisele Andrade, I greatly appreciate for your care, support and encouragement to help me go through the frustrations and difficulties. Special thanks to my English teacher and life mentor, Mrs. Lori Petro and her husband Mr. Wesley Petro, for your love, care, support, guidance in every aspect of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Cryphonectria parasitica causes chestnut blight.

Fungus *C. parasitica*

In 1906, William A. Murrill first reported a new fatal epidemic disease, chestnut blight, of native American chestnut (*Castanea dentata*), which he observed in New York City, New Jersey, Maryland, the District of Columbia as well as Virginia [1; 2]. The pure culture of the fungus causing the chestnut blight was obtained and isolated from the infected chestnut twigs collected in the New York Botanical Garden [1]. The disease starts from the fungus entering through a wound or dead limb of the tree and growing under the cortex in the layers of the inner bark and cambium. The symptom of the infection in chestnut tree bark is a brown and soft canker with numerous yellowish-brown fruiting pustules called pycnidia containing massive amounts of spores[1]. Both the asexual conidia and the sexual ascospores are able to be dispersed into fresh wounds of trees rapidly by wind and rain as well as transmitted by beetles and birds (Figure 1.1) [2; 3]. In time, depending on the size of the infected area, the fungus starts to form a canker around the infected spot, eventually girdling the stem or branch, leading to death [2]. Unfortunately, the mycelium of the fungus growing on the tree was found so active and well protected that no treatment was proved to be effective [1].

In the Murrill's report, the fungus was named *Diaporthales parasitica*, a sac fungi species based on the observation of the mycelium and spore morphology in the life cycle. In 1978, based on the DNA sequence comparison and phylogenetic analysis, the chestnut blight fungus was identified as one species of phylum ascomycetes, class Sordariomycetes, order Diaporthales, and family Cryphonectriaceae and revised with a worldly accepted new name *Cryphonectria parasitica* (Murr.) [4].

C. parasitica devastated the American chestnut *Castanea dentata*

American chestnut trees (*Castanea dentata*) were once dominant, the major hardwood trees in the forests of the eastern regions of the USA [5]. In the southern part of its range, American chestnut trees generally grew to 37 meters tall and 1.5 meters in diameter. Due to the wood being extremely hard and resistant to decay, the American chestnut was widely utilized for construction, furniture, railroad ties, musical instruments, and many other items [5]. The nuts used to be an important source of food for wildlife, domestic livestock, and humans. Furthermore, the tannins extracted from the bark and wood used to provide for the basis of a large tanning industry [4-6]. Since *C. parasitica* was introduced into North American and spread rapidly, it devastated the American chestnut throughout its natural range by killing about 3.6 million hectares of American chestnut trees, an estimated loss of four billion stems, within 50 years (Figure 1.2) [5; 7-8]. This filamentous fungus colonized in the wounded cambium of tree stems regardless of the size, except the root systems that are protected in the soil. Root-collar sprouts formed from the uninfected root systems allow the tree to start the asexual reproduction in the natural forests, but frequently become infected with blight fungus again [8], thus providing a continuous inoculum to the surrounding forest. Unfortunately, the rare

existence of the sexual reproduction in nature leads to the functional extinction of the American chestnut trees in the current forests [8].

Hypoviruses, a potential way to save American chestnut trees

C. parasitica was also reported in Europe in 1938, near Genova, Italy where the fungus spread rapidly and caused the serious cankers on European chestnut trees as well [9]. However, the damage caused by the fungus was found to have less severity in Europe than in North America, leading to the discovery of the double strands RNA virus, Hypovirus [9]. In 1951, some healed cankers with the fungal mycelium only growing on the outer layers of the bark was discovered resulting in the chestnut trees survival from the infection and this phenomenon was called hypovirulence [10]. Later, the hypovirulent fungal strain was found to contain the high molecular weight double-stranded RNA (dsRNA) leading to the white hypovirulence phenotype (Figure 1.3) [11]. Molecular characterizations have shown hypovirus lacks capsid structures [12]. The best characterized type of hypoviruses, the CHV-1 family, have dsRNA approximately 12.7 kb in length that contains two open reading frames (ORFs), ORF A encoding a papain-like protease that generates two polypeptides p29 and p40, and ORF B encoding a similar protease p48 at amino-terminus (N-terminal) and containing regions of RNA-dependent RNA polymerase (RdRp) and helicase (Hel) binding motifs, but with as yet undetermined mature protein products [13-15]. It was found that the dsRNA was successfully transmitted to the virulent strains by converting them to hypovirulence through hyphal anastomosis [16]. Since then, this chestnut blight fungus and its associated hypovirus have been widely studied to provide a potential biocontrol strategy to restore the chestnut trees from the blight in North America [9; 17].

Vegetative Incompatibility, limiting the transmission of hypoviruses among *C. parasitica* strains

In the United States, since 1975, the hypoviruses have been released into the natural forests for the biological control of chestnut blight [18]. However, almost all these efforts have resulted in the failure of transmitting the hypoviruses among the various strains of *C. parasitica* [18]. The main factor that has received most attention is the vegetative incompatibility system that restricts the hypovirus transmission between individuals. Vegetative incompatibility is a self/nonself-recognition system that results in the program cell death when cells of two incompatible individuals anastomose [19]. As a defense mechanism of fungi, vegetative incompatibility is genetically regulated to trigger programmed cell death to prevent the mixing of cytoplasm of one individual with another during hyphal fusion when the horizontal transmission of hypoviruses can occur [20].

Vegetative incompatibility is a common phenomenon in the filamentous ascomycete fungi. In model species *Neurospora crassa* and *Podospora anserina*, a few genes that control this process were characterized at the molecular level [21]. Heterokaryon incompatibility (*het*) genes of the two species were characterized to encode the proteins containing the conserved HET domain, which induces the recognition signal and activates the programmed cell death process [22]. In *C. parasitica*, vegetative incompatibility is controlled by at least six unlinked vegetative incompatibility loci (*vic* loci), with two alleles at each locus [23]. Individuals are compatible if they share the same alleles at all *vic* loci, but are incompatible if they differ at one or more locus. The six unlinked *vic* loci (*vic1, vic2, vic3, vic4, vic6, vic7*) have been characterized through linkage mapping and comparative genomics [24]. Most of them contains the common HET domain and NACHT/WD40 domain, which were found to associate with the

activation of programmed cell death in vegetative incompatibility [20; 25]. At *vic4,* two alleles encoding a 359 amino acids protein kinase c-like and a 1,628 amino acids NACHT/P-loop and WD40 repeats domains, respectively, were identified (*vic4-1* and *vic4-2*) [24]. The two individuals differing at the *vic4* (with *vic4-1* and *vic4-2* genotype respectively) were found to form the barrage triggered by the above genes during anastomosis, but it was found not to impede the transmission of the hypovirus like the other loci, and the mechanism of this phenomenon is still unclear [26].

CPVIB-1, a transcription factor that is essential for the vegetative incompatibility trigged by *vic4*

Transcription factors are a group of proteins that bind to specific DNA sequences to mediate the expression levels of target gene in response to a signal [27]. There is a class of p53-like transcription factors that is known as NDT80/PhoG-like DNA-binding family, the Ndt80 protein from *Saccharomyces cerevisiae* is one of the first members found to bind to the middle sporulation element (MSE) to trigger the expression of 150 genes during meiosis [28-29]. The number of NDT80-like genes in fungi are varied as well as their roles, which include regulation of meiosis, sexual development, virulence and the response to nutrient stress and programmed cell death [30-31].

In *N. crassa*, a mutant in an ORF named VIB-1 (vegetative incompatibility blocked) was found to suppress the phenotypic aspects of vegetative incompatibility trigged by the *het-c* locus [32]. VIB-1 was identified as a putative 670 amino acids transcriptional factor due to its NDT80/PhoG-like coding domain, an ortholog of Ndt80 gene of *S. cerevisiae* [33]. Later, VIB-1 was characterized to localize and aggregate in the nuclei supporting its DNA binding characteristics [33]. qPCR analysis revealed that VIB-

1 is essential for the expression of several genes involved in programmed cell death triggered by vegetative incompatibility, such as *pin-c, het-6,* and *tol* [32-33]. Further study suggested that CPVIB-1 functions as a regulator of hyphal compartmentation, death rates and conidiation in vegetative incompatibility as well as a repressor of glucose metabolism [34].

In the chestnut blight fungus *C. parasitica*, an ortholog of VIB-1 of *N. crassa,* the putative CPVIB-1 was identified by Rong Mu (unpublished study). CPVIB-1 was found to produce a protein of 649 amino acids in length, share 48% identity with VIB-1 of *N. crassa*, and contain a NDT80/PhoG-like domain (unpublished study). In order to characterize the role of the CPVIB-1 such as vegetative incompatibility, virulence, hyphal growth, and the programmed cell death process, a *cpvib-1* deletion strain was generated. The *cpvib-1* deletion strain was observed with phenotype shifting in various aspects, including virulence, hyphal growth, sporulation, and vegetative incompatibility regulation.

The Δ*cpvib-1* mutant strain shows reduced hyphal extension and profuse conidiation, indicating CPVIB-1 is required for fungal cell growth and to regulate the signals controlling sporulation (Figure 1.5). The deletion of *cpvib-1* largely decreasing the canker formation on the chestnut stems indicates that CPVIB-1 plays a direct or indirect role in virulence (Figure 1.6 (A)). Statistical analysis shows the canker size of the Δ*cpvib-1* strain is significantly smaller compared to both EP155 wild type (WT) strain and its hypovirus-infected strain (Figure 1.6 (B)). Also, the deletion of *cpvib-1* alters the barrage formation pattern between EU1 and EP155 strain, which differ from each other in the *vic4* loci (Figure 1.7). The barrage formation was observed between EU1 and EP155

(Figure 1.7 (A)) but disappeared when *cpvib-1* was disrupted from both sides (Figure 1.7 (B)) implying that CPVIB-1 is required for the barrage formation in vegetative incompatibility triggered by *vic4* (unpublished study, R. Mu).

In conclusion, consistent with the roles of VIB-1 in *N. crassa*, CPVIB-1 in *C. parasitica* is crucial for vegetative incompatibility triggered by at least one allelic mismatch, hyphal growth, sporulation, and pathogenesis.

The goal of this project

The primary objective of this study is to identify the direct and indirect targets of CPVIB-1 in *C. parasitica* to further build the complete pathway for the vegetative incompatibility system. As a transcription factor with a NDT80/PhoG like domain, CPVIB-1is hypothesized to bind specifc DNA sites in the genome to activate or repress the transcription of its targeted genes, which then affect their downstream factors. We propose two strategies to achieve the goal. The first strategy is the large scale transcriptome sequencing (RNA-Seq) of the EP155 wild type and its isogenic *Δcpvib-1* strain. Transcriptome comparison analysis is the critical step to provide the informative insights into the downstream effectors of CPVIB-1 with the well accepted bioinformatics pipeline and reliable genome reference. The second, more challenging, stategy is to identify the exact binding locations of CPVIB-1 by using ChIP-Seq (Chromatin Immunoprecipitation and Sequencing). This will use a tagged CPVIB-1 protein containing the FLAG peptide, cross-link it with its bound DNA, fragment the DNA, then immunoprecipitate the protein-DNA complex with specific antibodies to recover the DNA fragments bound to the CPVIB-1 protein, prior to sequencing the bound sites using a high throughput platform. A critical and initially overlooked factor that affects both the

transcriptome comparison analysis and ChIP-Seq analysis was the availability of an accurate and reliable genome reference and annotation, which brings up the second objective of this study.

The second objective of this study became a re-annotation of the *C. parasitica* genome with improved accuracy and enriched structual and functional information. Using outdated bioinformatics tools, the shortage of transcripts evidence (ESTs from Sanger sequences), and the discovery of mistakenly annotated genes (Willyerd, et al. unpublished data), the prior genome annotation (2009-version) from JGI (Joint Genome Institute) was hypothesized to be in a poor condition with misannotated gene models structurally and functionally. The general existence of the inaccurately predicted gene models could impact studies to identify specific functional genes as well as studies of revealing the regulation mechanisms using genome scale strategies. We therefore set out to re-annotate the genome of *C. parasitica* with the advanced MAKER2-two-pass pipeline to generate gene models to train the predictors and provide the quality metrics system, the RNA-Seq data to provide sufficient transcripts evidence, and the newly UniProt/SwissProt protein database to provide more experimentally reviewed protein evidence. With this re-annotated version (2017-version), the first objective of this study was carried out relying on this improved annotation to provide more accurate and informative results. However, with both the prior annotation and the re-annotated version available, it was necessary to carry out a comparison between them to report their differences in structural and functional predictions and their accuracy as well, which became the third objective of this study.

The third objective of this study was to develop a pipeline to evaluate the prior annotation of *C. parasitica* and compare it to the re-annotated version with the purpose of evaluating each of the gene predictions, and finally sort them into four categories corresponding to their accuracy and consistency between the annotations.

In summary, this study began with investigations of the vegetative incompatibility pathway of *C. parasitica* using newly available high throughput transcriptome and ChIP-Seq technologies. Fulfillment of those goals also necessitated re-annotating and validating gene model predictions in order to be confident that the results of our studies would be as accurate as possible. As a beneficial by-product of this work, we have developed a novel pipeline for genome annotation comparisons that is broadly applicable to other experimental systems, and also greatly improved the accuracy of genome resources available to the community of researchers working with *C. parasitica*.

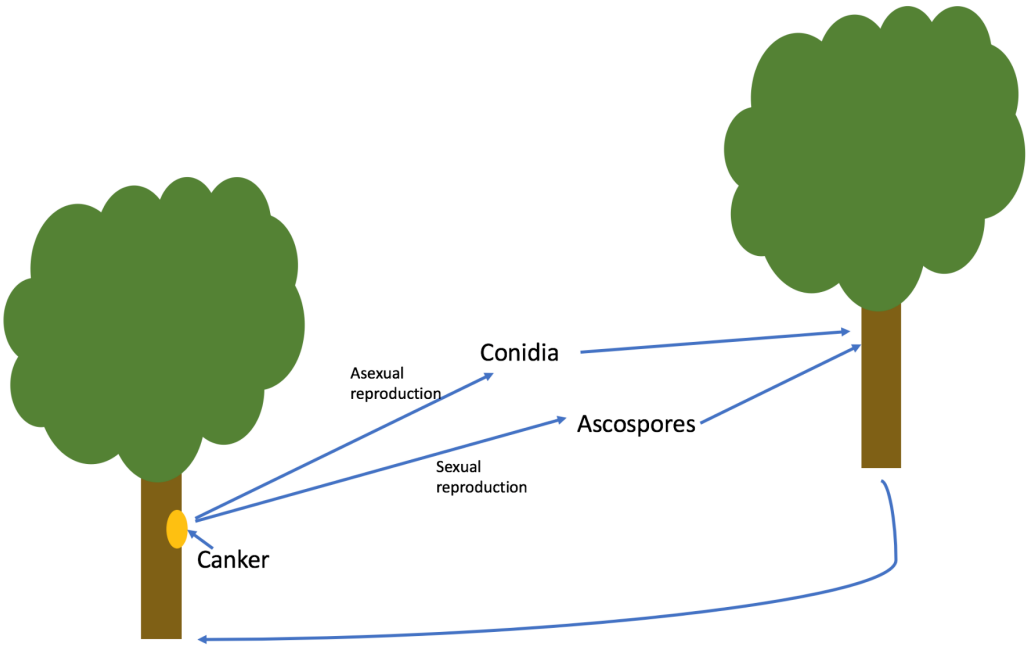


Figure 1.1 Infection cycle of *C. parasitica* in American chestnut trees based on information read in reference [3].

The mycelium in the infected canker region is able to produce conidia with asexual reproduction process as well as ascospores with sexual reproduction process by mating with spores from other strains. Both conidia and ascospores are able to enter fresh wounds in another American chestnut tree by the dispersion of vectors, such as wind, rain, and insects and germinate to orange mycelium leading to the formation of the canker in this infected regions.

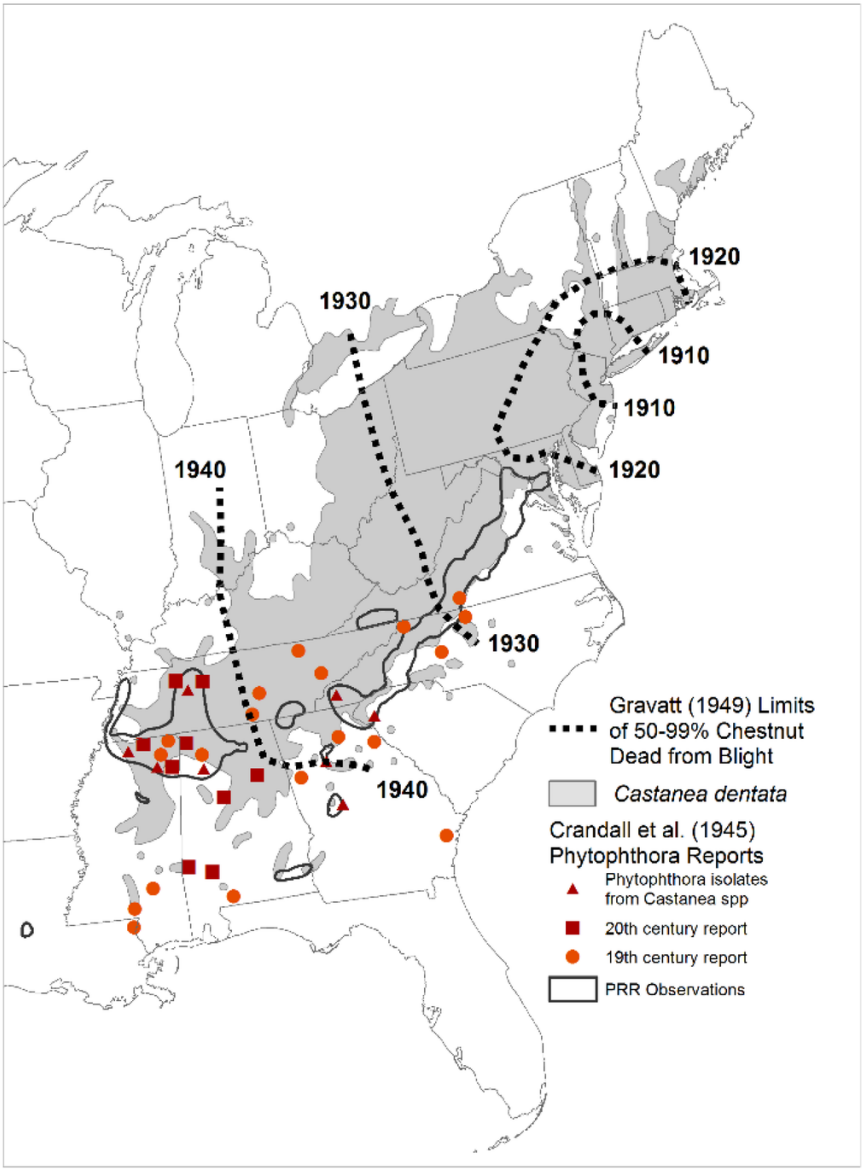
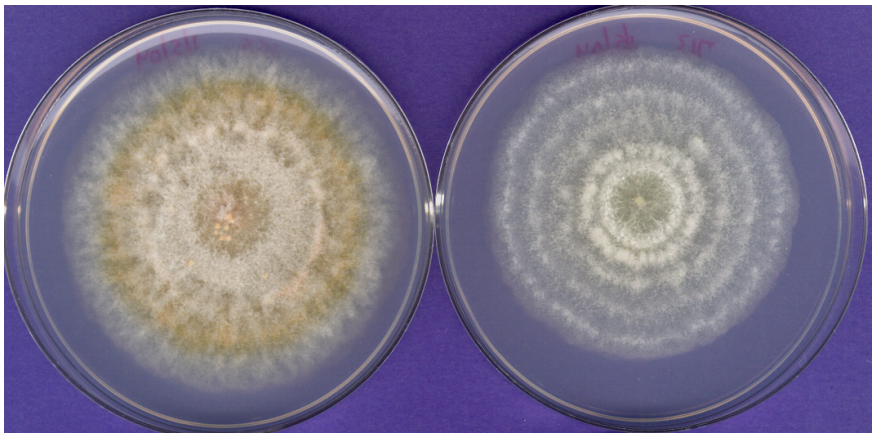


Figure 1.2 The natural range of American chestnut trees [8].

The gray shaded area represents the natural range of American chestnut populations; the dash lines represent the progress of the disease during the first half of the 20th century.



(A) (B)

Figure 1.3 The phenotype comparison of the two representative *C. parasitica* strains.

(A) represents the phenotype of EP155 wild type strain without hypovirus infection, (B) represents the phenotype of EP155 wild type strain with hypovirus infection.

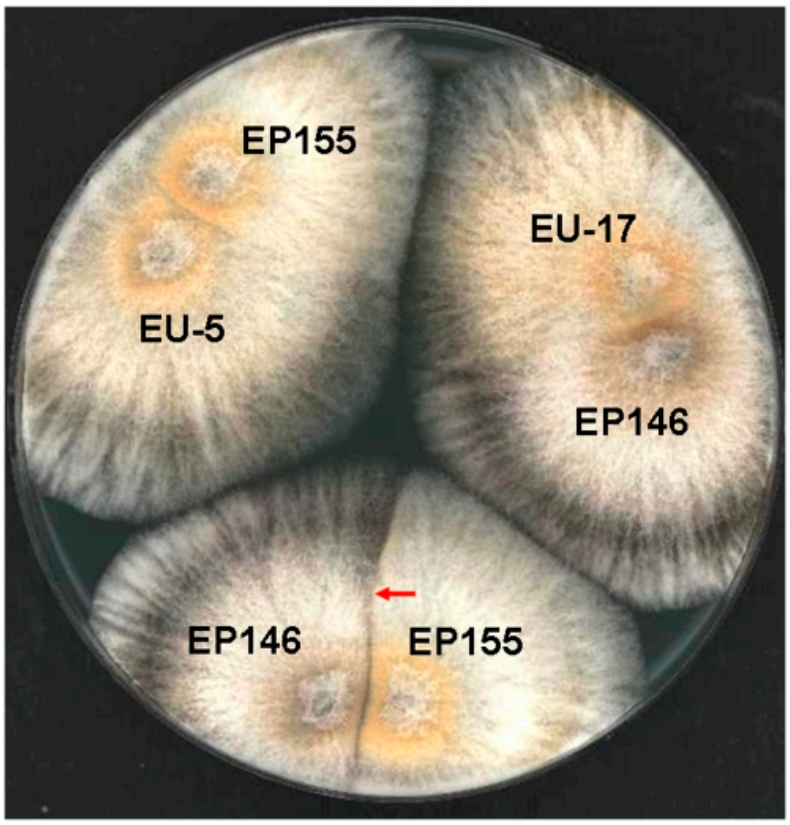


Figure 1.4 Vegetative incompatibility (*vic*) assay displaying the compatible and incompatible phenotypes [24].

The bottom image represents the incompatible phenotype with a red arrow pointing to the barrage lines (a line of dead cells) that forms when two individuals that differ at one or more *vic* locus merge and trigger the programmed cell death.
The left and right two images represents the compatible phenotype.

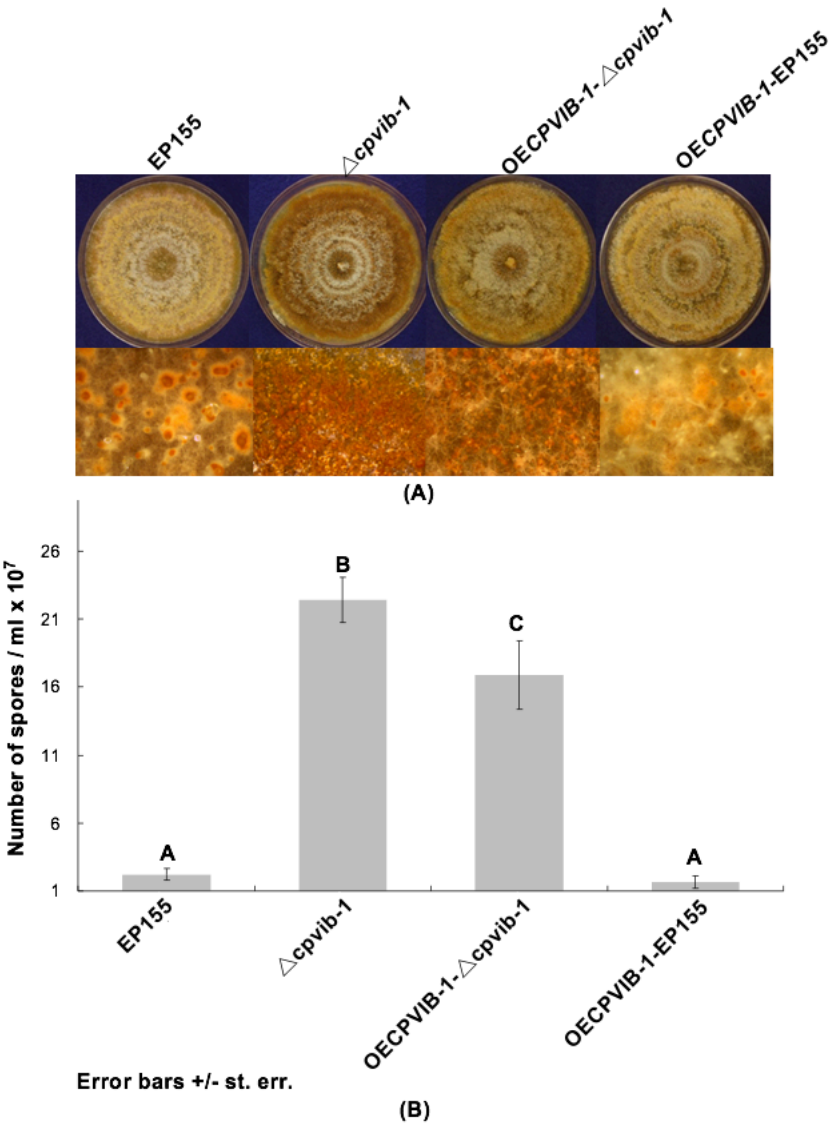


Figure 1.5 CPVIB-1 regulates sporulation and aerial hyphal growth [35].

(A) Morphological phenotype representatives of the *cpvib-1* deletion, complementary and overexpression mutants *(*Δ*cpvib-1*, OECPVIB-1-Δ*cpvib-1*, OECPVIB-1-EP155) as compared to the EP155 wild type strain when grown on petri plates. Sporulation and aerial hyphal growth patterns were shown under each strain. Δ*cpvib-1* strain produces profuse spores and reduced aerial hyphal growth compared to EP155 strain. (B) Quantitative analysis of sporulation using individual median test to determine pairwise differences with three biological replicates. Levels not connected by the same letter are significantly different [35].

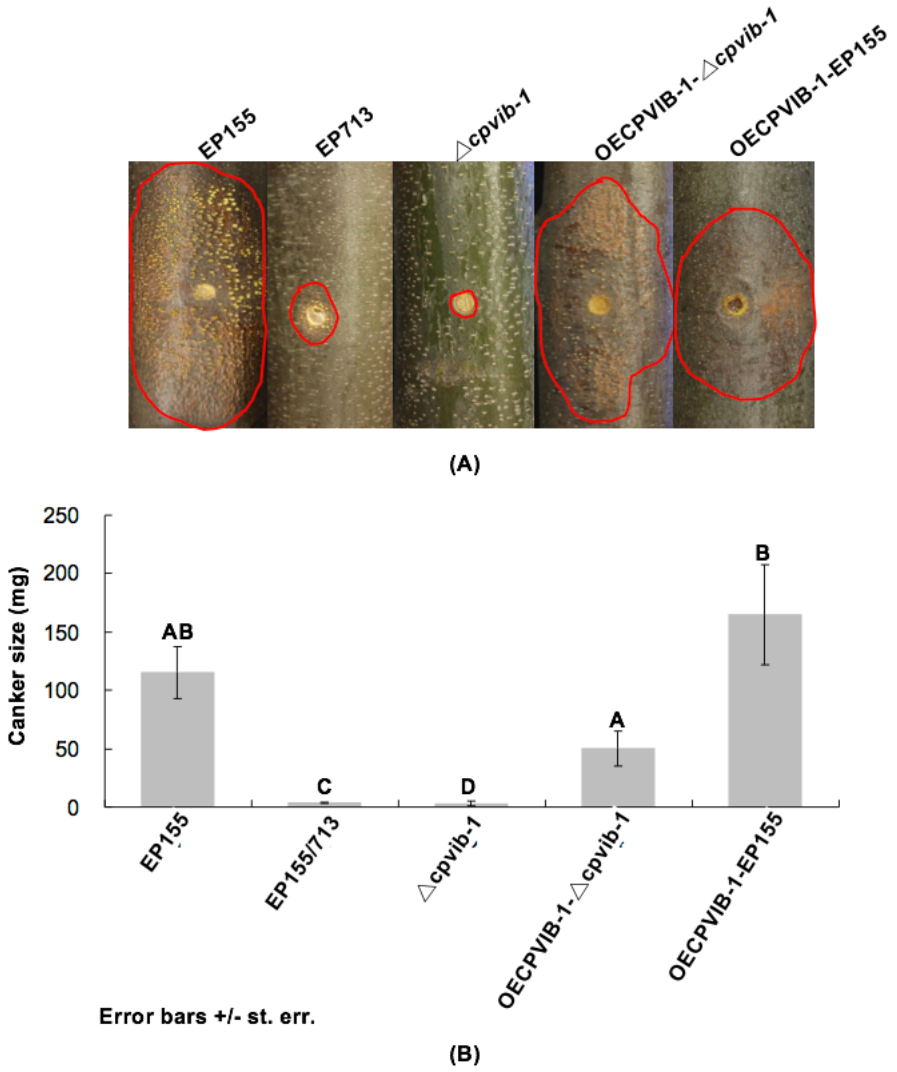


Figure 1.6 CPVIB-1 is important for fungal pathogenicity[35].

(A) Representative virulence assays on dormant chestnut stems carried out for 28 days. Canker sizes were labeled with red lines for each strain. (B) Graphical representation of the virulence assay using individual median test to determine pairwise differences with three biological replicates. Levels not connected by the same letter are significantly different. Y-axis: Canker size (mg) after 28 days [35].

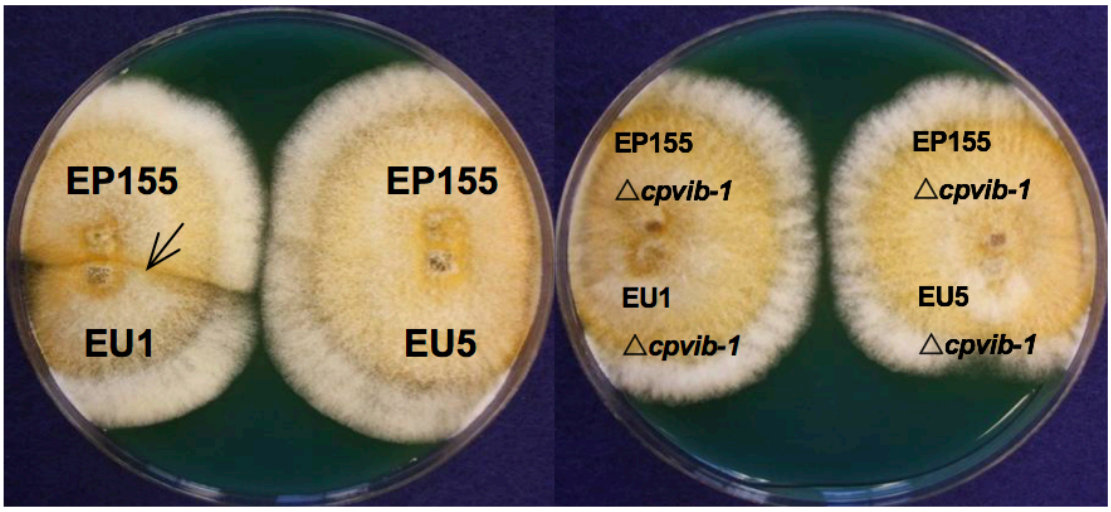


Figure 1.7 CPVIB-1 is involved in vegetative incompatibility [35].

(A) EP155 (WT) demonstrating incompatibility with EU1 but compatibility to EU5. (B) Deletion of *cpvib-1* from both EP155 and EU1 and EU5 changes the compatibility between EP155 and EU1 but keeps the compatibility unchanged between EP155 and EU5 [35].

CHAPTER II

RE-ANNOTATION OF THE GENOME OF CRYPHONECTRIA PARASITICA

Abstract

Next generation sequencing (NGS) technologies create great opportunities for various genomics, transcriptomics and proteomics projects of exotic non-model organisms. A well-annotated genome reference is critical for these projects like exploring target genes' functions or characterizing gene expression profiles. However, the errors in the annotation are often present in existing databases such as NCBI and JGI because of the lack of supporting evidence. Now, for exotic organisms, having the assembled genomes, detailed transcript evidence from RNA-Seq and more recent curated proteins database available, provides opportunities to perform a better quality new genome annotation or an improved re-annotation. Therefore, a practical MAKER2-two-pass pipeline was designed to perform a re-annotation of the genome of *C. parasitica*. This resulted in 11,171 predicted gene models, with 92% of them having coding domains matching various databases, and 78.36% of them having at least one conserved domain and important coding sites with an InterPro ID. The re-annotation provided new structural and functional features including, mRNA, UTRs and quality metrics system and ortholog proteins names, conserved domain IDs, and GO terms (Gene Ontology terms), separately. Together, these results suggested that the re-annotation provides an improved annotation version in both accuracy and information details.

Introduction

Next generation sequencing (NGS) techniques have dramatically dropped the costs of genome sequencing projects leading to unprecedented opportunities for many non-model organisms' genome projects [36]. At the time of writing, 5,386 eukaryotes genome sequences are recorded in NCBI genomes database with "scaffold or contigs" and "Chromosomes" status, indicating those are at the amenable stage for gene and genome annotation.

Most recent gene predictors for annotation were built based on the Hidden Markov Model (HMM), which defines probability distributions statistically for sections of genomic sequences in eukaryotes, like introns, exons, untranslated regions (UTRs), etc. [37]. The accuracy of gene prediction with the above tools mainly depends on the amount of evidence to pinpoint boundaries between each two states, such as intron-exon boundaries or UTR-exon boundaries[38].

Yet, in many ways, the genome annotation for non-model organisms has encountered a lot of challenges. One of the factors responsible for this is the absence of sufficiently large and reliable experimental evidence for the gene finder algorithms. Second is the lack of pre-existing gene models like the first generation of genome projects with relatively large resources [39].

By investigating gene expression profiles using RNA-Seq data aligned to a reference genome [40], the advance of NGS technologies and the corresponding bioinformatics tools have allowed more and more non-model organisms to have unprecedented opportunities for their unique metabolic pathways to be

explored. However, there are many challenges in obtaining the reliable results from such projects especially given the lack of a high-quality genome annotation.

The genome project of the *C. parasitica* was carried out in 2009 using predominantly Sanger sequencing technology, and along with its first version of gene predictions, was released by the Joint Genome Institute (JGI). Although this was a significant advance at the time, the bioinformatics tools used, and the lack of quality metrics, have made this information outdated. Contributing to this problem were the use of low coverage expressed sequence tags (ESTs), a then current (but now outdated) proteins database, and the fact that there were only 15 gene models of *C. parasitica* reviewed in Swiss-Prot database with experimental evidence. Considering these factors, it is no surprise that there are mistakenly annotated gene models structurally and functionally in the first annotation version of *C. parasitica*.

While it is not feasible for any genome annotation, to be 100 % accurate, inaccurately predicted gene models have the potential to impact the studies of functional genes and domains of interest that relate to all aspects of the organism's biology. Before the work performed for this study, there were at least four genes on different scaffolds of the *C. parasitica* Ep155 assembled genome that were identified as not matching the predictions from the first annotation (2009-version). Three of them were transcribed in different regions and one was found with different reading frame (Willyerd, Pokharel, Ren and Dawe, unpublished observations).

Furthermore, our initial transcriptome profiling project using knock-out strains for different genes of interest failed to provide efficient and informative output. Moreover,

these errors found in the *C. parasitica* Ep155 strain are common for a genome project of its age with insufficient evidence, annotation tools and strategy [41-42].

Fortunately, the NGS high-throughput RNA-Seq technique offers an enormous amount of assembled transcript data to help predict new genes and transcripts [41-42]. Out of all forms of evidence, high quality RNA-Seq assembled transcriptome sequences have the greatest potential to improve the accuracy of gene annotations, as these data provide copious evidence for better delimitation of exons, splice sites and alternatively spliced exons [39]. Compared to first-generation sequencing, NGS could generate more than 500 gigabases in a single run creating a great opportunity to improve annotation quality and reveal each transcript with high coverage, high sensitivity and high dynamic range [43].

In this study, we proposed to re-annotate the genome of *C. parasitica* taking advantage of more transcript evidence from the NGS transcriptome sequencing (RNA-Seq) and the newly updated protein evidence from the UniProt/SwissProt database consisting of all manually annotated and reviewed proteins. The Illumina HiSeq2500 platform was used to sequence three samples of the transcriptome from the *C. parasitica* Ep155 strain in the 100bp pair end mode. Each sample library yielded between 45 and 51 million 100 bp reads for a total of about 21 Giga bases of sequence data. With the abundant RNA-Seq data, the genome reference based transcriptome assembly was in order to provide assembled transcript evidence to the re-annotation process using the widely accepted Tophat and Cufflinks Tuxedo suite pipeline [44].

The second advantage of this study was to use the advanced MAKER2-two-pass optimum annotation pipeline, which is a configurable genome annotation and curation

pipeline incorporating three gene predictors, GeneMark-ES, SNAP and Augustus[38; 45-48].

An important feature of this pipeline is GeneMark-ES-3.0, a self-training gene predictor employing an unsupervised training procedure to produce a reliable *ab initio* gene model training sets for subsequent gene finder methods, SNAP and Augustus [47]. Then, the HMM based gene finder Augustus was used to find an optimal parse of a given genomic sequence. The advantage of using SNAP is to allow for both change of the underlying HMM and flexible inputs from the above steps, resulting in iteratively improved training sets with transcriptome and protein evidence [45; 49-50]. MAKER2 is an annotation management tool that combines all three of the above gene predictors along with built-in tools, including RepeatMasker, BLAST+, and Exonerate. It provides an easy-to-use way to either perform a *de novo* genome annotation with a new genome, or update a pre-existing annotation with quality-control metrics, by aligning protein and RNA evidence in a splice-aware manner to accurately identify splice sites [45].

The resulting *C. parasitica* genome re-annotation provides more accurate and informative predicted gene models than the prior annotation (2009-version). New structural and functional features as well as quality metrics system for each predicted gene model were generated in this study based on more evidence and better tools. Overall, this new annotation, named 2017-version, provided an improved annotation file and facilitates the future gene identification and characterization in *C. parasitica.*

Materials and Methods

Fungal transcriptome preparation and RNA-sequencing

Mycelia of *C. parasitica*'s wild type strain EP155 were cultured in potato dextrose broth (PDB; Difco, Sparks, MD) on the benchtop under ambient conditions. The cultures were homogenized and diluted with fresh media (1:1 volume) and allowed to grow for an additional 16-18 hours to achieve log phase fungal cells for RNA extraction. Following harvesting by filtration through Miracloth, the cultures were then immediately ground to a fine powder by using a sterilized mortar and pestle with liquid nitrogen [51]. Total RNA was then isolated from about 30 mg powered mycelia using RNeasy® mini kit (Qiagen) according to the manufacturers protocol. Next, the quality of the total RNA was determined by agarose gel and analysis on an Agilent 2100 Bioanalyzer® (www.chem.agilent.com). Briefly, 1 μg total RNA with a RNA integrity number (RIN) of 8 or higher was enriched for mRNA using Illumina TruSeq RNA sample preparation Kit v2, Set B (Illumina, #RS-122-2001) according to manufacturer's instruction. Next, the Bioanalyzer was used to detect size distribution of the transcriptome after fragmentation and adapter ligation, and a Qubit® Fluorometer (InvitrogenTM, Waltham, MA) was used to quantify the yield post-PCR amplification. Finally, the pooled cDNA libraries with 10 μM concentration from each sample were sequenced on Illumina HiSeq2500 platform in a 2 x 100bp paired-end (PE) configuration with High Output mode (V4 chemistry) at Genewiz (South Plainfield, NJ).

Reference based transcriptome assembly

A python pipeline script was written to perform one-step transcriptome assembly with the raw RNA-Seq FASTQ files obtained from three biological replicate samples

above by using the open source software programs of the Tuxedo suite and other auxiliary programs (Figure 2.1) [52]. First, a widely applied quality assessment software FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to exam the quality of the reads, which indicated no need for any trimming process. All sequencing files were then aligned to *C. parasitica* reference genome from JGI MycoCosm genome portal (http://genome.jgi.doe.gov/Crypa2/Crypa2.download.html) with Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) and TopHat (https://ccb.jhu.edu/software/tophat/index.shtml) to discover transcripts splice sites, using default parameters [44]. Subsequently, Cufflinks was utilized to assemble the transcripts from all successfully aligned reads of each sample, followed with running Cuffmerge to merge their assemblies together into an integrated reference transcriptome annotation file [42; 44; 52]. The python script file Transcritpome.assemble.pipeline.py is available from the GitHub website (https://github.com/didiren/transcriptome_assembly) and can be applied to run one-step transcriptome assembly in any system with Bowtie2, TopHat, Cufflinks and Python available [42; 52]. A README.md file has been included to demonstrate the usage of this pipeline starting with the raw RNA-Seq FASTQ files. The output generated was a file called Assembled_transcriptome.fasta that was then used for genome annotation in the next section.

MAKER2-two-pass genome annotation pipeline

The MAKER2-two-pass genome annotation strategy used in this study required three input files for this pipeline: the assembled *C. parasitica* genome file; the newly generated assembled transcriptome file; and, the well-characterized protein dataset uniprot_sprot.fasta downloaded from Swiss-Prot database [53-55]. These components

provided the intrinsic RNA evidence and the extrinsic protein evidence to improve the gene model prediction (Figure 2.2).

RepeatMasker was used to identify the repeats regions in conjunction with the compacted repeat libraries RepBase in MAKER2 package by taking all libraries as resource (Figure 2.2) [39]. Subsequently, GeneMark-ES-3.0, a self-training gene finder that can identify protein-coding genes, was run to creates the *ab initio* HMM gene models from the genome file only [56]. Next, the gene model predicted above was applied to the MAKER2-first pass annotation process to train its installed gene-predictors, SNAP and AUGUSTUS, as well as experimental evidence of exon-intron boundaries from transcriptome assembly and the protein dataset to improve the accuracy of gene prediction [23; 38; 46; 57; 58]. Because of the iterative fashion of MAKER2, the gene prediction process can achieve higher accuracies by running the MAKER2 a second time using the gene models built from the first pass as training data [45; 48; 59]. At the end of the second run, MAKER2 consolidated the output to a final annotation in GFF3 format (called 2017-version) by picking the post-processed gene models that were most consistent with the experimental evidence such as RNA and proteins alignments [39; 45; 59]. Along with the annotation file, the gene, transcript, CDS and protein sequences for the predicted final gene model were extracted in FASTA format. Although the structural predictions of the gene model are complete at this point, it is essential to assign putative gene functions to newly annotated genes. In order to build these annotations, the reviewed proteins dataset from UniProt/Swiss-Prot was used as reference to append the matching protein's name to each predicted gene model in the GFF3 file using BLASTp. Additionally, InterPro conserved domain ID and GO ID (https://www.ebi.ac.uk/interpro/)

were appended to the predicted gene models using InterProScan [60-62]. The detail for the above operations are illustrated in https://github.com/didiren/PEPA.

Results

Description of *C. parasitica* genome

The first step towards the successful annotation of any genome is a good genome assembly [39]. The Ep155 strain of *C. parasitica* used in this study is the laboratory-standard well-characterized, virulent strain (ATCC 38755) [63]. The latest genome assembly was completed with 26 main genome scaffolds in a size of 43.9 Mb by constructing the whole genome shotgun reads in 2010. Assembly parameters are reported by JGI as follows: Scaffold N50 is 5,118,729 bp, the estimated gap percent is 0.2%, and the coverage of genome is 99.6 % (http://genome.jgi.doe.gov/Crypa2/Crypa2.download.html).

Quality assessment of the RNA-Seq reads

The transcriptomes of three biological replicates of Ep155 strain generated 147.7 million reads for three libraries with the average mean quality scores of 35.57 to 35.71 and >=Q30 scores of 94.07 % to 94.53 % (Table 2.1). Generally, sequence runs with >50 % of the nucleotides having quality scores >Q30 are considered acceptable [64]. Therefore, 35.65 as the mean quality score of the three libraries combined, and >94% bases per read with a quality score >Q30, indicated the high quality of the sequences generated for the *C. parasitica* transcriptome.

Quality assessment of the transcriptome assembly

94.2 % to 95.1 % of the reads from the three samples were mapped to the reference genome, with 93 % of the reads being concordant from paired files of one sample (Table 2.2). This compares favorably with expected sequencing accuracy, given that 70 – 90 % of RNA-Seq reads from humans samples are expected to map onto the humans genome [65]. In order to improve annotation with the new transcripts evidence, the transcriptome assembly was performed without using the prior annotation as reference to reveal novel transcripts [65]. There were 12,926 to 13,037 novel transcripts in each transcriptome as defined by Cufflinks in General Transfer Format (GTF) format, and Cuffmerge united three of them to 14,707 novel transcripts for *C. parasitica* genome (Table 2.2).

New structural and functional features in the re-annotated version

Compared to the prior annotation file from JGI additional gene structures, such as mRNA, exon, coding sequences (CDS), 3'UTR and 5'UTR, were listed for each predicted gene. Also, quality metrics, including MAKER mRNA Quality Index (QI) tags and the Annotation Edit Distance (AED) scores, were added to each predicted gene model in light of new transcript and protein evidence. AED is a measure of the agreement between each predicted gene model and its supporting evidence, with a number between 0 and 1. An AED score of 0 denotes a perfect prediction with clear supporting evidence, while a value of 1 indicates a complete absence of evidence supporting the annotated gene model [59]. In addition, the predicted protein's name from UniProt/SwissProt dataset as well as the conserved domain information, such as InterPro ID, Pfam ID and GO ID, were also appended (Table 2.3).

Optimum predicted gene models from the MAKER2-two-pass pipeline

There were 8,278 gene models predicted in MAKER2-first pass preliminary annotation step and 11,171 genes models after the completion of MAKER2-second pass final annotation while using the full strategy (Table 2.4) (Figure 2.2). In contrast, only 7,419 and 7,272 predicted gene models were discovered without applying the *ab initio* gene predictor GeneMark-ES (Table 2.4) (Figure 2.2).

A previous study reported that a well annotated proteome would mean that 55 to 65 % of the predicted proteins should contain a recognizable domain [45; 48]. For these 11,171 predicted gene models, 10,285 (92 %) of them have hits from at least one domain database (such as CDD, Gene3D, Hamap, PATHER, Pfam, SMART), and of those 8,059 (78.36 %) have clear conserved domains and important coding sites with the InterPro IDs and GO IDs. This provides additional supporting evidence as to the effectiveness and accuracy of this new annotation.

Discussion

The accuracy and completeness of a genome annotation directly impacts the validity of individual gene or genetics-based functional studies [66]. During a study to identify potential virulence effectors and downstream factors of related transcriptional regulators in *C. parasitica*, the prior annotation (2009-version) resulted in miss-steps due to the mistakenly predicted gene models. It is especially critical for plant pathology, for which extensive research efforts have been invested into identifying and characterizing putative virulence-associated effectors to contribute to the control of disease and the protection of the plants against pathogens [67].

As mentioned above, the first step towards a successful genome annotation is to conduct an assessment of its genome assembly to describe its completeness and continuity with a N50 score, which is the most widely used statistics for describing the genome assembly quality [39]. Therefore, the N50 scaffold length of the *C. parasitica* genome equals to 5,118,729 bp indicates a high-quality genome assembly. Two other informative statistical parameters reflecting the quality of genome assembly are the percentage of gaps and the coverage of genome [39]. In this case, these values are 0.2 % and 99.6 %, respectively. Thus, the *C. parasitica* genome assembly obtained from Sanger sequencing was demonstrated to be a remarkably high-quality product.

When re-annotating this genome, the transcriptome assembly was generated from approximately 146 million 100bp mRNA reads, which have a mean quality score of above 35 (25 is considered decent score in NGS). Moreover, more than 94% of reads were aligned to the genome assembly leading to more than 300-fold increased coverage of the transcriptome in *C. parasitica*. The Tuxedo suite pipeline generated a large set of transcripts models, many of them overlapping one another. The final transcriptome assembly comprised 14,707 transcripts. This represented about 3,000 more novel transcripts and about 650bp longer average transcript length compared to the prior annotation (2009-verion).

In addition to new transcript evidence, the protein evidence was used in this study is from UniProtKB/SwissProt database, which has been highly recommended as an excellent core source of curated proteins [53]. This is the recommended database to use when working with an organism that has a limited number of reviewed proteins (such as *C. parasitica)* [39].

In the prior annotation, the bioinformatics tools were only capable of finding the most-likely CDS of a gene but did not report UTRs or alternative spliced variants. The MAKER2-two-pass pipeline for the genome annotation provides a much more complex output than simple gene prediction. It not only reports the CDS, exon and gene structural features based on the heterogeneous evidence, but also synthesizes gene models and produces an output that describes detailed features such as 5'UTRs, 3'UTRs, exon and mRNA, which can be visualized in genome browsers and annotation databases. Secondly, MAKER2 is the first bioinformatics tool to provide a quality metrics system. MAKER2 mRNA Quality index (QI) tags and the Annotation Edit Distance (AED) scores are included in the annotation process. Thus, the quality assessment provided a means to highlight any problematic predicted gene models for further manual curation and also gave a measure for the comparison of two annotation versions. The third feature in the re-annotation version is the functional addition to each predicted gene model of a conserved domain ID, GO ID and protein name from the UniProtKB/SwissProt database. With the addition of the three new features to the annotation it provided both more accurate predicted gene models and increased confidence in the information presented.

The overall quality assessment of an annotation has been suggested to consider three factors: the number of predicted gene models; the protein domain content; and, the AED scores. In this study, as a non-model organism, *C. parasitica* does not have a pre-existing gene model to train the gene predictors. The MAKER2-two-pass pipeline took the advantage of the specific self-training gene predictor, GeneMark-ES to provide a training gene model and iteratively improve the prediction quality resulting in 11,171 predicted gene models in the 43.9 million bp genome. As a reference, gene numbers from

the model fungus, *N. crassa*, has about 10,000 protein coding genes in approximately 40 million bp size genome (http://fungi.ensembl.org/Neurospora_crassa/Info/Index). Therefore, the total gene number in the re-annotation is similar compared to other related organisms. Secondly, 92 % of the predicted gene models (2017-version) have at least one coding domain, more than the 55 – 65 % usually considered to represent a high-quality annotation [59]. The third factor, AED scores, will be described in detail in Chapter 3.

In conclusion, the genome of *C. parasitica* was sequenced, annotated, and released in 2009 using the technology and tools available at that time. However, with deep transcriptome sequencing and updated protein evidence, re-annotation is now more accurate and informative. This improved genome annotation provides a valuable resource for researchers who are interested in both comparative and functional studies of *C. parasitica*. The integrated annotation analysis by applying the MAKER2-two-pass pipeline has facilitated the improvement of the genome annotation and this approach can be applied to other biological systems.

Figures and Tables

Table 2.1 RNA-Seq read quality from three biological replicates of the *C. parasitica* Ep155 strain.

Sample	Reads Yield (Million)	*% of >= Q30 Bases	Mean Quality Score
EP155s1	51.94	94.37	35.57
EP155s2	45.44	94.53	35.68
EP155s3	50.34	94.07	35.71

*% of >= Q30 Bases means the percentage of all baes in a read containing no more than one error in each 1000 bases [68; 75].

Table 2.2 The quality of read alignments and transcriptome assembly.

Sample	Read Yield (Million)	Read number mapped to genome	[a]Concordant pair alignment rate	[b]Discordant alignments rate	Transcript Counts (Cufflinks)	Total TranscriptCounts (Cuffmerge)
EP155s1	51.94	49.88 M (96.1%)	94.2%	0.3%	13,037	14,707
EP155s2	45.44	43.86 M (96.5%)	94.7%	0.3%	12,926	
EP155s3	50.34	48.92 M (97.2%)	95.1%	0.5%	13,010	

[a] Concordant pair alignment rate, referring to the percentage of reads from both paired-end sequencing file matched to the same locus in the genome.
[b] Discordant alignment rate, referring to the percentage of reads from both paired-end sequencing file, did not match to the same locus in the genome.

Table 2.3 New structural and functional features in the re-annotated version compared to the prior annotation.

	Prior version	Re-annotated version
Structural features	gene	gene
	exon	mRNA
	CDS	exon
		CDS
		three_prime_UTR
		five_prime UTR
Functional features	None *	Ortholog proteins name from UniProt/SwissProt
		InterPro domain ID
		GO ID
		Extra domain databases ID

* indicates there are no functional features attached in the prior GTF annotation file, but the JGI website has protein names, domain ID and GO ID for each predicted gene model if you examine each individual gene model.

Table 2.4 Comparison of the predicted gene model when applying the different strategy of MAKER2-two-pass pipeline.

	Predicted gene models number	
	Use the first pass[a]	Use two-pass[a, b]
With GeneMark-ES ([c]Red)	8,278	11,171
Without GeneMark-ES ([c]Red)	7,491	7,272

[a] referred to the steps highlighted in blue in MAEKR2 two-pass pipeline. [b] referred to the process highlighted in green in MAEKR2 two-pass pipeline. [c] referred to the process highlighted in red in MAEKR2 two-pass pipeline.

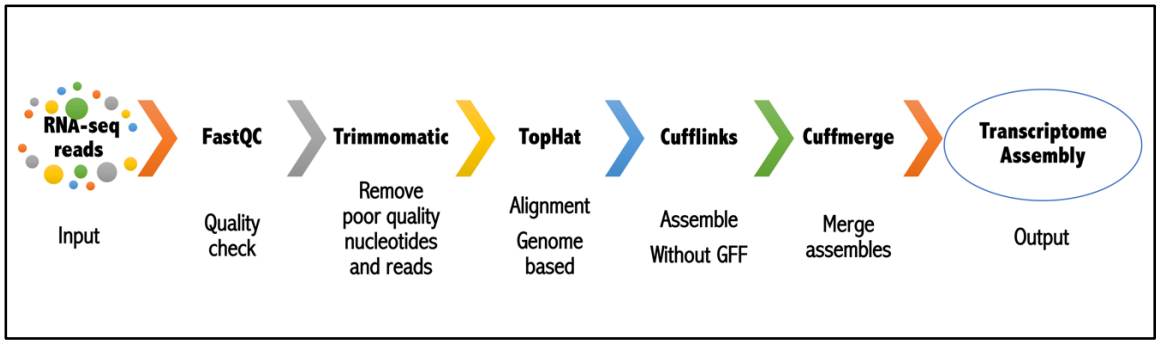


Figure 2.1 Transcriptome assembly Tuxedo suite pipeline strategy.

All RNA-Seq fastq files were fed into this Tuxedo suite pipeline with the beginning step of quality examination by FastQC. Trimming process was optional based on the quality of the reads from the former step. Then, the alignment was performed against the reference genome using TopHat, the transcripts were assembled individually with Cufflinks and merged into one integrated GTF file with Cuffmerge.

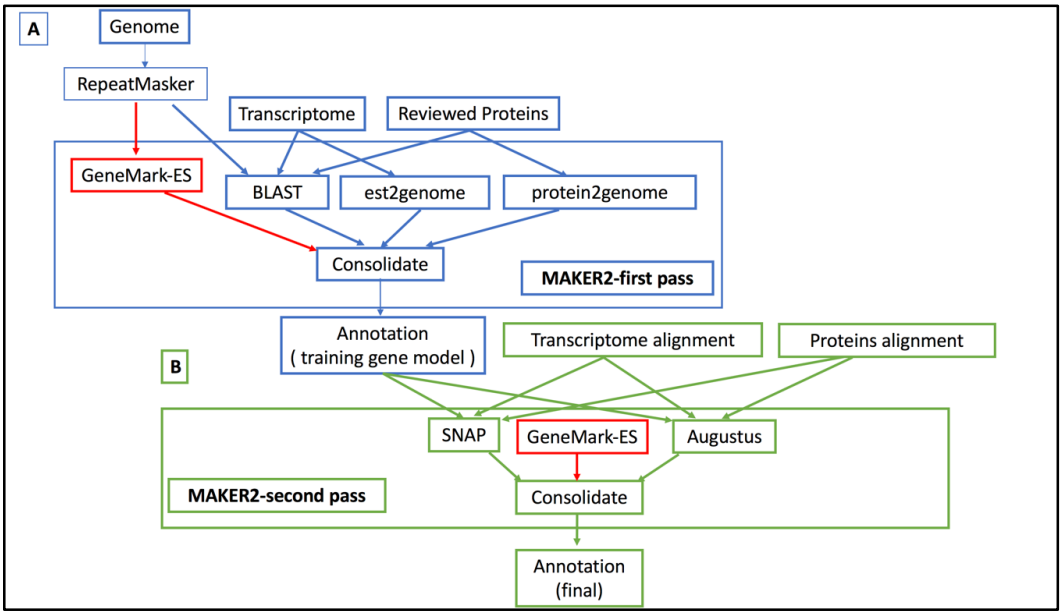


Figure 2.2 MAKER2-two-pass genome annotation strategy.

(A) The strategy started from the prerequisite RepeatMasker included in MAKER2, then in the MAKER2-first pass, an external gene predictor GeneMark-ES highlighted in red were used to provide a training HMM file along with the RNA and proteins evidence for the second pass gene predictors, showing a significant improvement in the number of predicted genes. (B) In the MAKER2-second pass, the gene models generated above were applied to train the other gene predictors, SNAP and Augustus along with MAKER2 internal programs to improve the gene models.

CHAPTER III

PEPA: A PIPELINE TO COMPREHENSIVELY EVALUATE A PRIOR GENOME ANNOTATION AGAINST A NEWER VERSION

Abstract

For the non-model organisms, a well-annotated genome reference is critical for either exploring single target gene function or characterizing whole genome expression profiles. Recently, some integrated genome annotation pipelines were developed to combine the evidence of abundant transcripts and the improved protein evidence to increase the performance of genome annotation. However, in terms of all target genes predicted and annotated from a prior and potentially outdated genome annotation, it is essential to evaluate the accuracy and discrepancy compared to the newer version. In the case of *C. parasitica* a simple pipeline, PEPA, was developed to comprehensively estimate the accuracy of each prior predicted gene model (2009-version) with updated transcript and protein evidence using the MAKER2 legacy annotation program, and thereby enrich the gene models with a quality metrics system and internal domain information (InterPro ID). Subsequently, a comparison of the prediction quality of all gene models from the prior annotation (2009-version) and the re-annotation (2017-version) from above was carried out. A python script was developed to sort each individual gene model from the prior genome annotation (2009-version) into four categories (Match, Similar, Different, Noexist based on their comparison to the newer

version (2017-version)). Twenty-seven of the predicted gene models comprising representatives from the Similar, Different and Noexist categories were then experimentally validated using a diagnostic PCR design to amplify the distinct regions to differentiate between the possible mRNA products predicted by the different annotation results. Of 11,609 predicted gene models from the 2009-version, only eight (0.06 %) were completely supported with transcript and SwissProt protein evidence.

When the two annotations were compared using the PEPA pipeline, 32.73% of predicted gene models were sorted into the Match category, 33.19% into the Similar category, 30.84% into the Different category and 1.1% into the Noexist category. Subsequently, 22 out of 27 chosen predicted gene models (from 2009-version) were then experimentally proven by PCR to support the re-annotation. Altogether, the results showed the general existence of errors in the 2009 annotation and the applicability of the PEPA evaluation pipeline in identifying the potential mistakenly predicted gene models to any genome annotation project.

Introduction

NGS encourages researchers to obtain the genome sequence of the exotic, non-model, organisms rapidly and with relatively low cost. Consequently, the accumulation of the genome annotations is rapidly increasing but they are mainly accomplished with automated annotation systems rather than something with more manual input as can be accomplished for model organisms supported by large research communities [69-70]. Many institutes, such as Joint Genome Institute [71] and University of Maryland's IGS (Institute for Genome Sciences) annotation engine (http://www.igs.umaryland.edu/research/bioinformatics/analysis/), offer annotation

systems for various genome projects of bacteria and fungi. However, a reality of using automated annotation systems is the existence of errors in the output annotation, which can mislead future work. Also, when annotated once, these projects may not be updated to take advantage of the most recent tools and databases available to improve the annotation.

For the same set of genome data different annotation systems may generate different results based on their various methods [72]. However, it is difficult to decide which one is more suitable and accurate without comparing them, since a universal standard is not available. Therefore, quality control is the significant issue for all genome annotation projects.

The definition of genome annotation generally refers to the structural and functional annotation [39]. The approach for testing the quality of the structural annotation is to establish the quality metrics that measure the consistency of each annotation with its overlapping evidence, such as protein and transcriptomic data [39]. One example of a quality metrics system that was developed by Sequence Ontology Projects is the Annotation Edit Distance (AED). This was applied in an integrated annotation pipeline, MAKER2, which is able to automatically calculate AED for each predicted gene model [39; 73].

Considering the impact of potentially incorrect gene model predictions on further studies of charactering individual target genes and also genome scale expression profiles, re-evaluating genome annotation projects that were prepared without any quality metric system is of considerable importance.

Here, PEPA uses MAKER2 to not only carry out a re-annotation utilizing new mRNA-Seq data and updated protein data to improve annotation quality, but also to add quality metrics like AED to a legacy-annotation [48].

Then, PEPA will evaluate the prior genome annotation, regardless of the format, against a newer version to illustrate the structural and functional differences of each gene model and its accuracy and sort each predicted gene model into four categories (Match, Similar, Different, Noexist) based on discrepancy between the structural (start/end coordinates) and functional (coding domain InterPro ID) information compared to the same predicted gene model from the newer version.

To test the efficacy of this approach, we used the most recently released genome annotation of *C. parasitica* from 2009, publicly available from the USDA Joint Genome Institute (JGI; http://genome.jgi.doe.gov/Crypa2/Crypa2.download.html). Both the quality metrics AED system and conserved coding domain InterPro ID were appended to this legacy data. Following this process for *C. parasitica*, we found that only approximately 1/3 of the original predicted genes were considered the same in both annotations. Furthermore, when tested individually by diagnostic PCR, 81.5 % of a selection of different annotations supported the newer version.

This evaluation pipeline of a prior genome annotation has major implications for future work by providing reliable insights of each predicted gene model with the new quality metrics and identifying which specific gene predictions may be problematic. Especially in non-model organisms, using more efficient and accurate results with reliable evidence to improve the structural and functional accuracy of a predicted gene

model is extremely important for future characterization by specific studies and also for large scale transcriptomics data analysis.

Materials and Methods

Transcriptome evidence and protein evidence

The same transcriptome and protein evidence were used here as described in the Materials and Methods of Chapter II.

MAKER2 legacy protocol

Like the MAKER2-two-pass pipeline described in the Materials and Methods of Chapter II, MAKER2 legacy annotation protocol provided the means for employing the transcript and protein evidence to train its installed gene finders to evaluate the accuracy of the prior genome annotation. The details for this operation are illustrated in the README.md file available in https://github.com/didiren/PEPA.

InterProScan protocol

The same tool and protocol was used here as described in the Materials and Methods of Chapter II. The commands for this operation are illustrated in https://github.com/didiren/PEPA.

Visualizing the quality distribution comparison of legacy annotation and re-annotation version

An R script was developed to visualize the qualities of the predicted gene models at the genome-level by displaying the cumulative distribution of AED scores [59]. This R script file AEDdistribution.R is available in https://github.com/didiren/PEPA**.**

Sorting genes predictions from the prior annotation by their discrepancies against the re-annotation

A custom python script was developed to sort each predicted gene model from the prior version to four categories, Match, Similar, Different, and Noexist, based on the discrepancies of coding regions (start/end coordinates) and coding domains (InterPro ID) against the newer version (Figure 3.1). The script ultimately generated a text file for each category with the gene model ID, the coordinates, the AED scores and the protein ortholog names from both annotation versions attached. If the predicted gene models from the prior annotation shared the exact same coordinates for their CDS with the newer version, they were sorted to the Match category (Figure 3.1). If their CDS coordinates were different but their transcript coordinates overlapped with the newer version gene models and they shared the same conserved InterPro domain ID, they were sorted to the Similar category (Figure 3.1). Thus, the Different category included those predicted gene models from the prior annotation that shared different conserved InterPro domain ID in addition to overlapped transcript coordinates with the newer version (Figure 3.1). The last category, Noexist, contained the predicted genes from the prior annotation that were not predicted to be present in the newer version as well as any that were not supported by any RNA evidence (Figure 3.1). This python scripts file named PEPA.py is available in my GitHub website (https://github.com/didiren/PEPA.git). The README.md file demonstrates the usage of this script starting with two annotation files as well as the output from the MAKER2 legacy and InterProScan protocol described above.

Validation of 27 predicted genes from the prior annotation showing discrepancies with the re-annotated prediction

There were 27 genes in total, seven from the Similar category, six from the Noexist category and 14 from the Different category chosen to experimentally validate the quality of the newer annotation. For each, the specific primers were designed based on the distinct region of their predicted transcript from the two annotation versions. Polymerase chain reaction (PCR) was used to amplify those regions from both cDNA and genomic DNA. All the products were viewed on a 1% (w/v) agarose gel to examine the accuracy of each prediction. Selected genes and their primers are listed in Appendix A chapter.

Results

The quality comparison of the predicted gene models from the prior and newer annotation version

As one of the two important factors to evaluate a genome annotation, the quality metrics AED scores cumulative distribution curves were displayed for each annotation set (Figure 3.2). Although 9,127 out of 11,609 (78.6%) predicted genes from the 2009 annotation set have an AED of less than 0.5, only 3,262 (28.10%) have AED scores less than 0.3 and eight predicted genes have a AED score of zero (Table 3.1). This implies that there are significant disagreements between the predicted exons or splice sites and the newly assembled transcriptome and protein evidence (Table 3.1).

In a second estimate of improved quality, we used the old data (EST evidence from JGI) with the newer annotation tools (the MAKER2-two-pass pipeline). In this case, the quality of the annotation set was improved with 402 (3.5%) out of 11,371 predicted genes with the AED score of zero, 6,180 (54.3%) with the AED score less than 0.5, and

5,003 (44.00%) with AED scores less than 0.3, which indicated a much lower prediction accuracy (Figure 3.2). Therefore, although there were incremental improvements, the older EST data was not fully effective at providing an accurate annotation even with newer bioinformatics tools (Table 3.1).

In contrast, the newer re-annotation set was generated with the assembled transcriptome and current protein evidence with the same MAKER2-two-pass pipeline. This returned a much-improved annotation build with 2,747 (24.6%) out of 11,171 predicted genes having the perfect AED score of zero, 7,741 (69.30%) having the AED score less than 0.3, and 9,392 (84.1%) having the AED score less than 0.5 (Figure 3.2; Table 3.1).

Sorting the predicted genes in categories

Using all 11,609 predicted genes from the prior annotation, 3,800 (32.73%) of were sorted into the Match category, meaning they share the same CDS coordinates and domain ID with the predicted genes from the newer (Figure 3.3). 3,854 (33.19%) of them were sorted into the Similar category and 3,581 (30.84%) of them were sorted into the Different category (Figure 3.3). 126 (1.1%) sorted into the Noexist category meaning they lacked any supporting evidence (Figure 3.3). Any remaining predicted genes from 2009 were not in any category because despite the existence of a transcript alignment there was no gene model prediction in the newer annotation (Figure 3.3). This may be because, for the *C. parasitica* genome, the protein evidence dataset in UniProtKB/SwissProt is derived from experimentally reviewed data of model organisms, so these may represent a set of genes that are specific to *C. parasitica*, or at least to its relatives more closely related than the nearest model organisms.

Quality comparison of the predicted gene models in the four categories of the compared annotations

Figure 3.4 shows a comparison of the AED scores in the predicted gene models from each category. The majority of predicted gene models in the Match, Similar and Different categories of the newer version have AED scores equaling to 0. In contrast, for the prior annotation there were two peaks for all three categories which contained the majority of predicted gene models, one is between the 0.4 and 0.5, the other one is 1.0 (Figure 3.4).

Out of 3,800 genes in the Match category, there were 2,491 genes from the prior annotation with a conserved domain InterPro ID, but the newer annotation had 599 more genes (3,090) with at least one conserved domain (Table 3.2). The possible reason for a different protein product from the same CDS sequence is the reading frame shift or the inaccuracy of translation in the JGI workflow with a given annotation file. No differences were found in the Similar category because the strategy required that they must share the same domain information (Table 3.2) only. There were 663 more predicted genes in the newer annotation coding for at least one conserved domain in Different category than the 1,058 predicted genes of the prior annotation (Table 3.2).

Validation of the new predicted gene models by PCR

From the amplified transcript bands in agarose gel pictures of Appendix A chapter, for the Similar category, six out of seven chosen predicted genes were proven to support the prediction of the newer version. For the Different category, 11 of 14 predicted genes supported the newer version, while five of six predicted genes in the Noexist did not produce detectable transcripts experimentally, which is consistent with the prediction of

the newer version (Table 3.3). In total, for the 27 predicted gene models of the prior annotation tested by PCR that were predicted to produce different transcript products in the new annotation, 81.48% supported the newer predictions (Table 3.3).

Discussion

This study has provided a tool (PEPA) that can help to evaluate the prior annotation of a genome by analyzing and comparing the accuracy of each predicted gene, as well as the structural and functional features of each predicted gene, against the newer version, with no limitation of the format and the source of the prior annotation of a genome.

To test the efficacy of this tool, the prior annotation of *C. parasitica* was used as a test. From the accuracy analysis and comparison, we have observed that the old annotation, although based on the best available tools and information at the time, contained many ambiguities and incorrect predictions. Even using the older evidence with newer tools did not compensate for the lack of coverage provided by EST data only. However, when used in conjunction with RNA-Seq data prepared by the MAKER2-two-pass pipeline used in Chapter II, the product was much more accurate gene models.

The PEPA tool presented here also provides the utility of sorting and comparing the output of the annotations. This is especially important in order to be able to pursue specific genes for functional studies. Given that 65.92 % of the prior predictions were sorted in Match or Similar categories, this suggested that approximately 2/3 of the predicted gene models have trustworthy annotation predictions in the older version. However, that also implies that approximately 1/3 of the prior predictions were incorrect,

with potential significant impact on further studies of either individual genes or large scale analyses.

In addition to analyzing computational metrics such as AED scores to evaluate the efficacy of the new annotation predictions, we also validated 27 specific genes experimentally using PCR primers designed to differentiate between the old and new predictions. The observation that 21 (81.48%) of them provided unambiguous support for the new annotation is further strong evidence of the improvement in accuracy. However, it also points to the important fact that with current available technology no annotation can be considered to complete or 100 % accurate.

Before this study, there were multiple tools designed to compare two annotations, such as BEACON and gffcompare, or semi-automated genome annotation comparison [74-75]. BEACON was limited and only suitable for GenBank format annotation file [74], while gffcompare was designed to compare two annotations in genome scale statistically that does not provide the individual gene-level of resolution as produced by PEPA. The semi-automated genome annotation comparison scheme was designed to perform the functional comparison only between two annotations [75]. PEPA was designed to comprehensively compare two annotations of one genome in both accuracy and structure/function. A limitation of PEPA is the requirement to run multiple tools, such as MAKER2, as well as and R and python scripts.

Although tested here on the *C. parasitica* genome, PEPA is broadly applicable to other experimental systems where new data is available that was not present when an original genome annotation were prepared. This makes it possible to help the users to have better confidence in the predicted genes they are interested in and provide data to

design diagnostic PCR primers to validate their own specific areas of interest, and to provide better support for future large-scale sequencing analyses.

Figures and Tables

Table 3.1 The distribution of the predicted gene models' AED scores in both genome annotations.

	Total Predicted gene number	Predicted gene number with		
		AED=0.0	AED≤0.3	AED≤0.5
2009-version	11,609	8 (0.06%)	3,262 (28.10%)	9,392 (80.90%)
2017-version	11,171	2,747 (24.60%)	7,741 (69.30%)	9,128 (81.71%)
Control-version	11,371	402 (3.55%)	5,003 (44.00%)	6,180 (54.35%)

Table 3.2 The distributions of the predicted gene models with conserved domain in each category from both annotation version.

	Total Predicted gene number	Gene number With domain	Match		Similar		Different	
			With Domain	Without Domain	Match Domain	Without Domain	Mismatch Domain	Without Domain
2009-version	11,609	8,309	2,491	1,309	3,288	566	1,058	2,523
2017-version	11,171	8,059	3,090	710	3,288	566	1,721	1,860

Table 3.3 Validation results of the 27 chosen predicted genes from the prior annotation and newer annotation.

Category	Predicted gene	Annotation	Gene ID	AED	Scaffold	Start	End	Predicted gene size (bp)	Predicted transcript PCR size (bp)	Positive in PCR	PCR supportive annotation
Similar	1	2009	258862	0.48	scaffold_5	920289	920820	531	400	yes	2017
		2017	Ep155_U_T00006215_1	0.25	scaffold_5	918882	921612	2730	2261	yes	
	2	2009	356517	1	scaffold_5	924032	924674	642	330	no	2017
		2017	Ep155_U_T00006216_1	1	scaffold_5	924078	924454	376	292	yes	
	3	2009	255909	1	scaffold_4	1076299	1077428	1129	998	yes	2009
		2017	Ep155_U_T00004868_1	0.1	scaffold_4	1076177	1077593	1416	1051	no	
	4	2009	231803	0.45	scaffold_3	4888579	4889682	1103	/	/	2017
		2017	Ep155_U_T00004505_1	0.11	scaffold_3	4885525	4890764	5239	2100	yes	
	5	2009	355955	0.22	scaffold_4	1069744	1070666	922	/	/	2017
		2017	Ep155_U_T00004864_1	0	scaffold_4	1069009	1070853	1844	817	yes	
	6	2009	263897	0.36	scaffold_7	1689695	1691232	1537	/	/	2017
		2017	Ep155_U_T00008600_1	0	scaffold_7	1688290	1692388	4098	153	yes	
	7	2009	58876	0.4	scaffold_1	678670	679423	753	/	/	2017
		2017	Ep155_U_T00000186_1	0.01	scaffold_1	678481	679500	1019	380	yes	

Table 3.3 (continued)

Category	Predicted gene	Annotation	Gene ID	AED	Scaffold	Start	End	Predicted gene size (bp)	Predicted transcript PCR size (bp)	Positive in PCR	PCR supportive annotation
Different	8	2009	98319	0.41	scaffold_5	3000738	3002855	2117	1068	yes	2017
		2017	Ep155_U_T00006769_1	0.06	scaffold_5	2993355	3001226	7871	1609	yes	
	9	2009	231853	0.45	scaffold_6	2410893	2412035	1142	1052	yes	2009
		2017	Ep155_U_T00007748_1	0.25	scaffold_6	2408261	2412803	4542	/	/	
	10	2009	102253	0.84	scaffold_5	4277195	4277888	693	2105	yes	2017
		2017	Ep155_U_T00007119_1	0.36	scaffold_5	4274791	4277389	2598	2303	yes	
	11	2009	357202	0.38	scaffold_6	2896210	2897553	1343	/	/	2017
		2017	Ep155_U_T00007829_1	0.05	scaffold_6	2895033	2897647	2614	320	yes	
	12	2009	261603	0.41	scaffold_6	2523249	2525371	2122	570	yes	2009
		2017	Ep155_U_T00007774_1	0	scaffold_6	2522966	2525960	2994	280	yes	
	13	2009	346358	0.36	scaffold_4	4656868	4660251	3383	550	no	2017
		2017	Ep155_U_T00005807_1	0.02	scaffold_4	4657590	4661908	4318	574	yes	
	14	2009	245160	0.38	scaffold_1	3176047	3177203	1156	/	/	2017
		2017	Ep155_U_T00000872_1	0.22	scaffold_1	3172638	3176405	3767	258	yes	
	15	2009	357444	0.43	scaffold_7	1259884	1260550	666	/	/	2017
		2017	Ep155_U_T00008473_1	0	scaffold_7	1258718	1261375	2657	1022	yes	
	16	2009	222652	0.45	scaffold_6	1371443	1372342	899	208	no	2017
		2017	Ep155_U_T00007540_1	0	scaffold_6	1369252	1373733	4481	145	yes	
	17	2009	346810	0.03	scaffold_5	3366779	3373770	6991	3027	yes	2009
		2017	Ep155_U_T00006866_1	0	scaffold_5	3366769	3374058	7289	873	yes	
	18	2009	260426	0.31	scaffold_5	570069	572567	2498	1919	no	2017
		2017	Ep155_U_T00006108_1	0.04	scaffold_5	570975	573657	2682	1080	yes	

Table 3.3 (Continued)

	19	2009	94097	0.03	scaffold_6	279282	280501	1219	110	no	2017
		2017	Ep155_U_T00007212_1	0.01	scaffold_6	279282	280496	1214	157	yes	
	20	2009	322230	0.35	scaffold_4	3221369	3222209	840	/	/	2017
		2017	Ep155_U_T00005405_1	0.01	scaffold_4	3220823	3222517	1694	972	yes	
	21	2009	231988	0.37	scaffold_1	7077089	7078251	1162	/	/	2017
		2017	Ep155_U_T00001901_1	0.05	scaffold_1	7074829	7078602	3773	2511	yes	
Noexist											2017
	22	2009	65946	1	scaffold_1	168020	168910	890	352	no	
											2009
	23	2009	241925	1	scaffold_1	2700456	2702059	1603	750	yes	
											2017
	24	2009	75444	1	scaffold_11	865322	866791	1469	1195	no	
											2017
	25	2009	334967	1	scaffold_10	685961	687176	1215	430	no	
											2017
	26	2009	71384	1	scaffold_2	1034513	1035594	1081	777	no	
											2017
	27	2009	68001	1	scaffold_4	1226218	1227846	1628	525	no	

For some genes, if both annotations have yes in ‘Positive in PCR’ column this indicated one of their amplifying regions was completely within of the second one. In that case ‘PCR supportive version’ listed the longer region version, which is the correct prediction. For the genes, ‘/’ represented no PCR process were necessary.

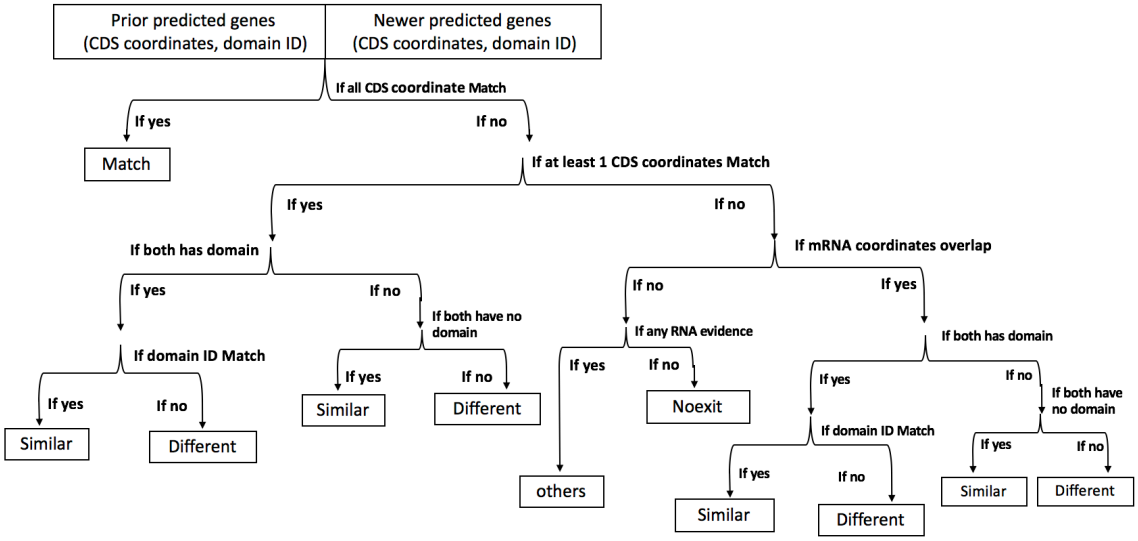


Figure 3.1 The newly developed custom Python sorting schematic diagram

Predicted genes from the prior annotation (2009-vesion) were sorted into four categories, Match, Similar, Different, and Noexist based on their structural and functional discrepancies with the newer set (2017-version).

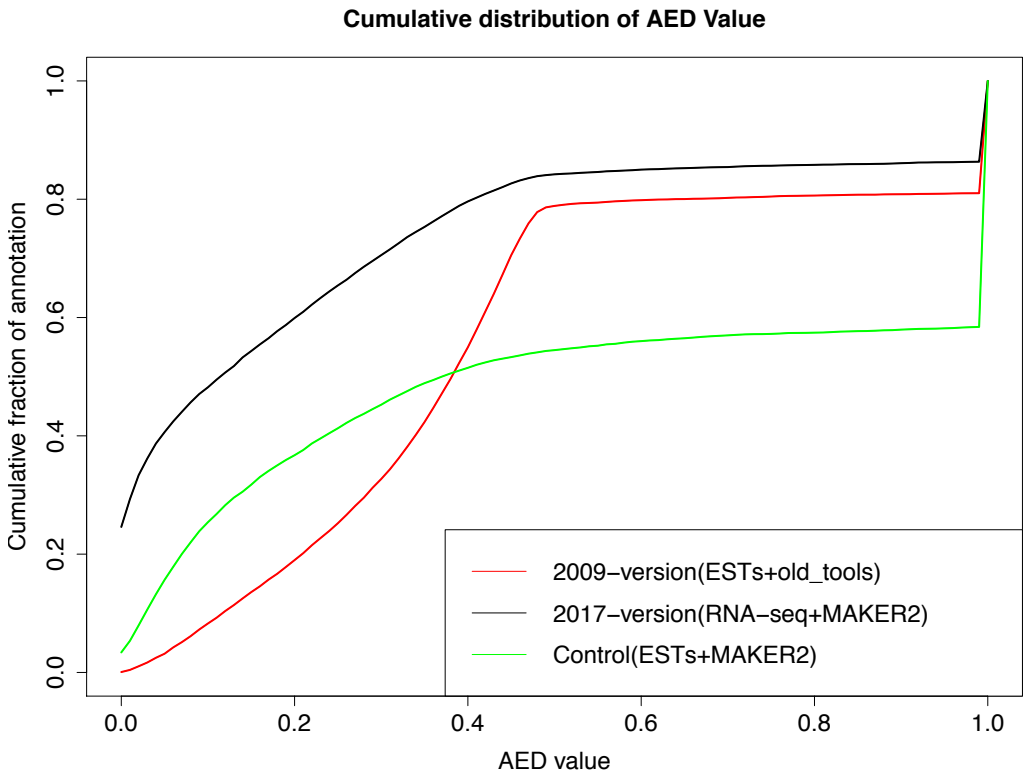


Figure 3.2 Cumulative distribution of AED values in both annotation versions.

Shown on the x-axis is the AED score from 0 to 1, and y-axis is the cumulative distribution of AED for each annotation version. (1**)** The red line presents the gene models from 2009-version annotation completed by JGI tools with only ESTs as RNA evidence. (2) The black line shows the gene models from 2017-version annotation generated with MAKER2-two-pass pipeline with RNA-Seq produced transcriptome as RNA evidence. **(3)** The green line is a gene models generated as pipeline control using MAKER2 tools and only ESTs.

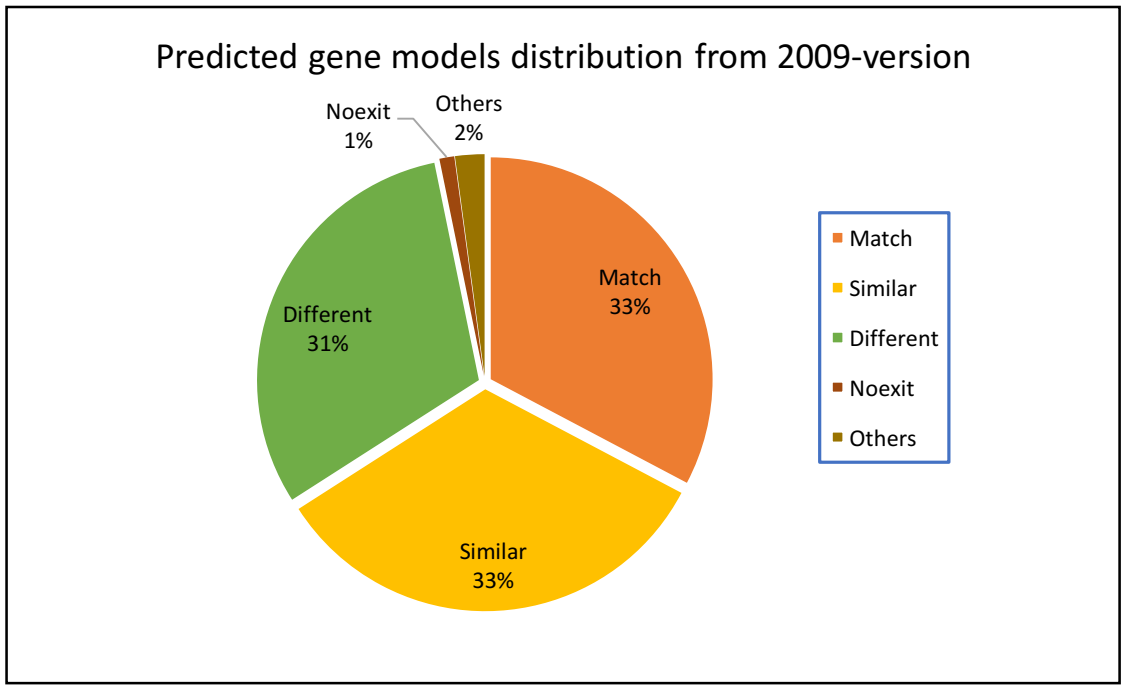


Figure 3.3 The distribution of predicted gene models from the prior annotation (2009-version) in the four categories.

(Orange pie) represents the portion of the total predicted gene models from the prior annotation that were sorted into the Match category. (Yellow pie) represents the portion of the total predicted gene models from the prior annotation that were sorted into the Similar category. (Green pie) represents the portion of the total predicted gene models from the prior annotation that were sorted into the Different category. (Maroon pie) represents the portion of the total predicted gene models from the prior annotation that were sorted into the Noexist category. (Brown pie) represents the portion of the total predicted gene models from the prior annotation that were sorted into the Others category.

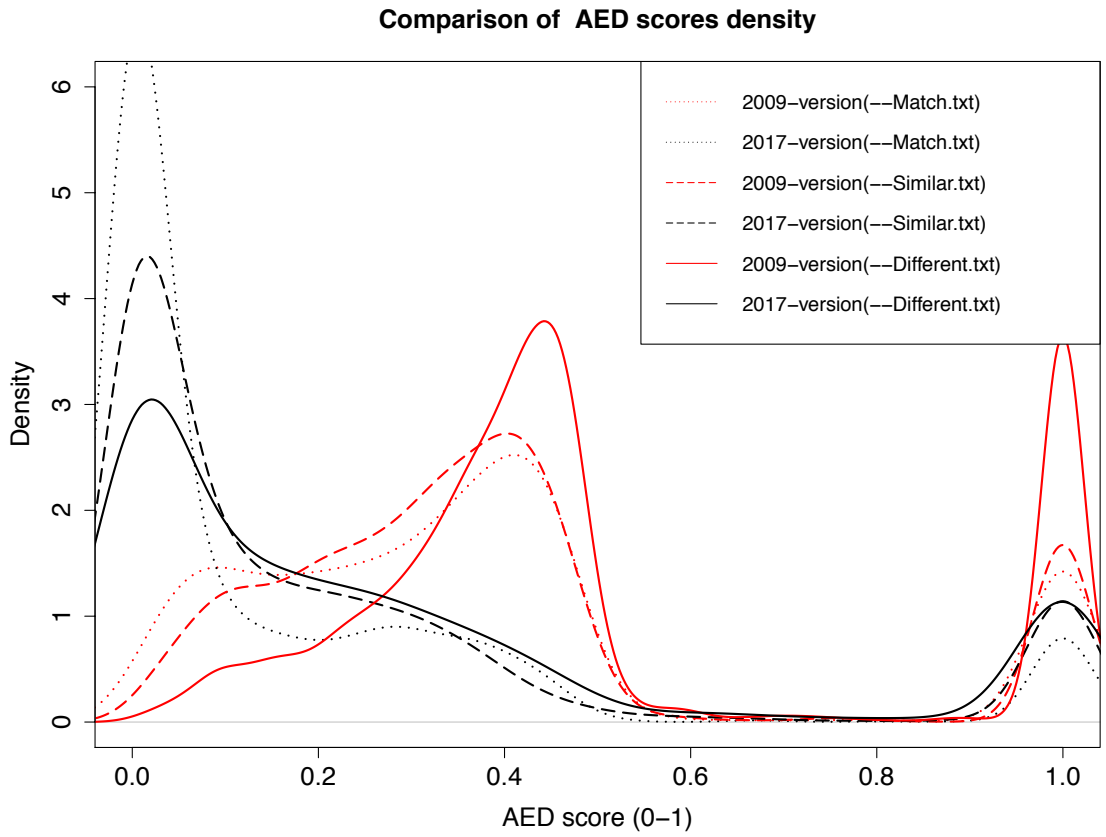


Figure 3.4 The AED score density curves of the predicted gene models from the Match, Similar and Different categories of both annotation sets.

(1) Red line presents the AED scores density from DIFFERENT category genes of 2009-version. (2) Red dash-line presents the AED scores density from SIMILAR category genes of 2009-version. (3) Red dot-line presents the AED scores density from MATCH category genes of 2009-version. (4) Black line presents the AED scores density from DIFFERENT category genes of 2017-version. (5) Black dash-line presents the AED scores density from SIMILAR category genes of 2017-version. (6) Black dot-line presents the AED scores density from MATCH category genes of 2017-version.

CHAPTER IV

EXPLORE THE DOWNSTREAM TARGETS OF CPVIB-1 USING TRANSCRIPTOME PROFILING OF THE CPVIB-1 MUTANT AND ITS WILD TYPE STRAIN

Abstract

C. parasitica is the causal agent of chestnut blight, which devastated the American Chestnut population in the early 20th century. The discovery of hypoviruses that reduce the severity of the chestnut blight infection offers the potential for biological control. However, the spread of the hypoviruses is hampered by a diverse nonself-recognition system, vegetative incompatibility (*vic*), among strains. VIB-1 in *N. crassa* has been reported as a transcriptional regulator that is required for the expression of downstream effectors of *N. crassa's* nonself-recognition system, heterokaryon incompatibility (*HI*). CPVIB-1, as an ortholog of the VIB-1, was identified as a transcription regulator playing an important role in programmed cell death in response to allelic variation at the *vic*4 locus. In order to explore the downstream targets that interact with CPVIB-1 to mediate the various phenotypic changes observed in the Δ*cpvib-1* mutant strain, including enhanced pigmentation and conidiation, less pathogenicity and reduced *vic* triggered program cell death, RNA-Seq was used to profile the variation of expression patterns between Δ*cpvib-1* mutant and wild-type strain. The high-quality RNA-Seq reads, were aligned against the *C. parasitica* genome using both the prior and newer genome annotation. After generating

a list of read counts per transcript to indicate the significantly changed transcripts a visualized expression pattern contrast between mutant and wild-type strain was used to examine biological processes altered by the absence of *cpvib-1*. In the absence of *cpvib-1*, there were 1,064 transcripts were altered significantly, with 245 (23%) up-regulated and 819 (77%) down-regulated. CPVIB-1 was found to be a key universal transcription activator required for preventing oxidative stress, for glucose signaling pathway, for DNA, protein and lipids synthesis, and for the pathogenesis.

Introduction

The pathogen that causes the chestnut blight, *C. parasitica,* was first discovered in New York City in 1904 [18]. After the introduction from the imported chestnut nursery stock, *C. parasitica* spread very rapidly, resulting in a devastating epidemic in American chestnut population. [5]. While *C. parasitica* almost caused complete destruction of *C. dentata* in USA, another chestnut species, *Castanea sativa*, suffered too in Europe. However, 10 years after the disease was first reported in Europe, the transmissible hypovirulence strains were discovered and identified to contain the transmissible RNA virus attributing to the striking reduction in fungal virulence [16]. Later the transmissible RNA virus was identified as a member of the hypovirus family and successfully applied as a biological way of controlling the spread of *C. parasitica* in Europe [76].

Unfortunately, the same strategy was not effective in North America [18]. The reason for the failure was later proposed to be vegetative incompatibility, which induces the sealing of the fused compartments that subsequently undergo programmed cell death between incompatible strains [77] and limit cytoplasmic exchange. Vegetative

incompatibility in *C. parasitica* is a nonself-recognition system, which is genetically controlled by at least six vegetative incompatibility (*vic*) loci [23; 63; 66; 77].

The mechanism of the genetic regulation of the vegetative incompatibility has been studied extensively in one of the ascomycete species, *N. crassa* to reveal a broad regulatory process [32]. As one of the studied targets, VIB-1 was discovered to be required for the programmed cell death and the hyphal compartmentation triggered by its heterokaryon incompatibility (*het-c*) locus [32; 63; 66; 78]. As a transcriptional activator with a DNA binding domain NDT80/PhoG-like, VIB-1 was found to be a regulator of conidiation, non-repressible acid phosphatase, other heterokaryon incompatibility locus (*het-e, het-6, het-8*) [63; 66; 78], downstream effectors associated with heterokaryon incompatibility, and the carbon and nitrogen metabolism [32-33; 79].

As for the chestnut blight pathogen *C. parasitica*, CPVIB-1, a putative ortholog of VIB-1 was identified coding a protein of 688 amino acids in length, sharing 40% identity with VIB-1 from *N. crassa* and containing the same NDT80/PhoG-like DNA binding domain. It was reasonable to hypothesize that CPVIB-1 might represent a potential candidate for modulation by hypoviruses by subverting the vegetative incompatibility. In order to test this hypothesis, a *cpvib-1* deletion strain (Δ*cpvib-1*) was established to test its role in vegetative incompatibility system. The disruption of *cpvib-1* in wild type EP155 strain exhibited reduced virulence, reduced aerial hyphal growth and profuse conidiation. Furthermore, CPVIB-1 was demonstrated to be required for programmed cell death and barrage formation between two incompatible strains with different alleles in *vic4* locus in the vegetative incompatibility assay. Thus, it is clear that CPVIB-1 plays a role in modulating the signaling pathways for at least one (*vic4*) allele interaction in *C.*

parasitica as well as the virulence, hyphal growth, pigmentation, and the conidiation process.

With the knowledge of VIB-1 mediating multiple downstream targets associated with heterokaryon incompatibility and the carbon and nitrogen metabolism in *N. crassa,* we proposed the hypothesis that CPVIB-1 is a mediator of the vegetative incompatibility, the virulence, and the fungal vegetative growth processes requiring the interactions with the downstream effectors.

In this chapter, the first strategy we proposed to achieve the goal of testing the above hypothesis was to carry out a comparative analysis of the large-scale transcriptome profile between the Δ*cpvib-1* strain and its isogenic wild type EP155 strain. RNA-Seq is a well-established high-throughput approach to obtain transcriptome profiling using the NGS deep-sequencing technologies, which provides a far more precise measurement of levels of transcripts and their isoforms than other methods [80].

In this study, three biological replicates of the above two strains were cultured in the same conditions and their transcriptome library was prepared and sequenced in the same lane using the RNA-Seq strategy described in chapters II and III. A widely applied transcriptome differential expression bioinformatics analysis pipeline was used to access the quality of the RNA-Seq reads, and align them to the reference genome, assemble the reads into transcripts, and count the read number mapped to each transcript [42; 81]. These data were used to infer functional and mechanistic pathway changes from the Δ*cpvib-1* strain.

Consistent with the proposed role of VIB-1 made in *N. crassa*, CPVIB-1 was found to be a universal transcription activator that regulates more than 1,000 genes in *C.*

parasitica. Moreover, compared to only the general metabolism pathways were highlighted to be significantly regulated using 2009-version annotation, there were 17 metabolism pathways were significantly up-regulated including six carbon metabolisms and 3 metabolism pathways down-regulated including RNA processing metabolism pathways in the Δ*cpvib-1* strain.

Materials and Methods

Fungal transcriptome preparation and RNA-sequencing

The wild type strain of EP155 was obtained from Dr. Donald L. Nuss (Center for Biosystems research, UMBI) and its isogenic *cpvib-1* mutant strain was obtained from Dr. Rong Mu (New Mexico State University).

The preparation of transcriptome and RNA-sequencing for the wild type EP155 strain of *C. parasitica* were demonstrated in Chapter II. At the same time and with the exactly same procedure, the transcriptome of the *cpvib-1* mutant strain was prepared and sequenced in the same Illumina lane.

Differential expression analysis of the EP155 strain and its isogenic Δ*cpvib-1* strain

Similar to the one-step transcriptome assembly python scripts demonstrated in Chapter II, a python script pipeline was developed here to perform the first segment (steps in blue frames) of the differential expression analysis workflow of the given two sets of RNA-Seq data, each from 3 biological replicates samples (Figure 4.1). The first step of the pipeline is to inspect the reads quality of each RNA-Seq raw sequencing file with FastQC [82]. Based on the quality output from the last step, Trimmomatic was the optional tool for trimming the adapters and poor quality reads if required [83]. Then, all

remaining paired-end reads were aligned to the *C. parasitica* reference genome from JGI, the same one used in the Chapter II and III, with Bowtie2 and TopHat [42; 44; 52; 84-85;]. Subsequently, all mapped reads were sorted with Samtools and counted per transcript with HTseq-count using both the 2009-version annotation and 2017-version annotation as references. [86-87].

The python scripts file differential_expression.pipeline.py can be downloaded from the GitHub website (https://github.com/didiren/RNA-seq-differential_expression) and run for one-step transcriptome assembly in any system with Bowtie2, TopHat, Samtools, HTseq-count and Python available. The details of how to run the pipeline are in the README.md file at the same web address.

Differential expression analysis with DESeq2

The reads count text files output from HTseq-count were fed into DESeq2, an R package providing the means to test differential expressed genes from the Δ*cpvib-1* strain compared to the EP155 strain by using the negative binomial generalized linear models [88]. In this study, it is the first step of the second segment (steps in orange frames) in the differential expression analysis workflow (Figure 4.1), included in a R script to summarize the reads count from all six samples to a condition dependent count matrices, analyze the count matrices for differentially expressed genes, visualize the results and cluster samples and genes using transformed counts. This R script is available in the GitHub website (https://github.com/didiren/RNA-seq-differential_expression) mentioned above.

KEGG pathways enrichment analysis with GAGE and Pathview R packages

As the second step of the second segment (steps in orange frames; Figure 4.1), a few tools were utilized together to discover the significantly regulated metabolism pathways in the Δ*cpvib-1* strain compared to the EP155 strain. At first, the BLASTp program from NCBI was used to match all genes from *C. parasitica* to their ortholog genes in *N. crassa*. Then the ortholog KEGG ID of *N. crassa* was extracted and replaced the original *C. parasitica* gene ID in the output reads count text files generated from HTseq-count. Second, to enrich sets of related genes sharing the same metabolic pathways from the gene expression data, the Generally Applicable Gene-set Enrichment (GAGE) was applied to this study [89]. All reads count text files with the *N. crassa* KEGG ID now were fed to the GAGE package to highlight the significant regulated metabolic pathways in the Δ*cpvib-1* strain compared to the wild type strain based on the integrated p-value from all related genes in the same pathways. The R/Bioconductor package, Pathview, was used to visualize the individual metabolic pathways from KEGG by automatically downloading the pathways graph data, parsing the data file, mapping the output data from GAGE to the data file and integrating the results to a new Graphviz graph along with the native KEGG view graph [90-91]. A python script and R script to run ID conversion and GAGE as well as Pathview are listed in the GitHub website (https://github.com/didiren/RNA-seq-differential_expression) mentioned above.

GO enrichment analysis with REVIGO

As the last segment (steps in green) of the differential expression analysis workflow (Figure 4.1), a web server, Reduce and visualize gene ontology (REVIGO) was used to summarize long, unintelligible lists of GO terms by finding a representative

subset of the terms using a simple clustering algorithm that relies on semantic similarity measures [92]. Essentially, the significantly differential expressed genes from DESeq2 that were statistically interpreted with the p-values lower than 0.05 and the log2 fold change larger than +2/-2 were subsequently linked to the GO ID of each gene by a small python script. Next, REVIGO was used to cluster and visualize the non-redundant GO term set to facilitate the interpretation of the output of differential expression of large set of genes from DESeq2 [92].

Results

Quality assessment of RNA-Seq reads and alignments in Δ*cpvib-1* strain

The transcriptome of three Δ*cpvib-1* strain biological replicates were sequenced with the isogenic wild type EP155 strain in the same lane using the Illumina paired end sequencing technology. There were 24 to 47 Million reads for these three libraries with the average mean quality scores of 35.57 to 35.75 and >=Q30 scores of 93.82% to 94.74% (Table 4.1). With the quality examination by FastQC, there were no adapters or poor quality reads with a mean quality score lower than 25 detected in all three Δ*cpvib-1* strain libraries, and the same for EP155 samples. In the light of the discovery that trimming process of RNA-Seq reads could alter the expression pattern estimation of samples, the Trimmomatic step was skipped in this study to avoid the unnecessary bias leading to the inconsistent results [93]. After the alignment with Bowtie2 and TopHat, there were 95.45% to 97.02% reads from the three samples of the Δ*cpvib-1* strain mapped to the reference genome along with approximately 93% reads being concordant from paired files of one sample (Table 4.2). The above results indicated that the quality of the

libraries for both the Δ*cpvib-1* strain and the EP155 strain was consistent among the biological replicates.

Differentially expressed genes in the Δ*cpvib-1* strain

The matrix with six columns and 11,171 rows (after mapping to 2017-version annotation reference) were normalized by modeling with negative binomial distribution with both the mean, dispersion values and the individual normalization size factors of each gene. Then, both the fold change and its dispersion (estimated standard error) between the treatment (Δ*cpvib-1*) and control (EP155) were included in the reduced process to report the significantly differentially expressed genes. In this case, the log2 fold change plot was a useful overview of the differential expression analysis of the Δ*cpvib-1* strain compared to EP155 strain results from DESeq2 (Figure 4.2). This plot demonstrated that some genes with small normalized count (in the left part of Figure 4.2) were also called significant with a small p-value, leading to the requirement of the DESeq2's shrinkage estimation strategy by considering the dispersion factors within the group variability. The red trend line showed the dispersions' dependence on the mean that was set to shrink each gene's estimates (black dots) towards to the trend to obtain the final estimates (blue dots) as well as the dispersion outliers (blue circles) were cut from the total gene sets because of the high gene-wise dispersion estimates (Figure 4.3). The dispersion value was estimated to be 0.01 in this study represented the gene's expression tends to differ 10% between samples of the same group [88].

It was important to perform the sample difference and clustering analysis between the two strains in this study to validate that the designed RNA-Seq experiments fit to our

expectation. In order to fulfill the goal, at first, the regularized-logarithm transformation (rlog) strategy of DESeq2 was used to get rid of the bias of a few low reads count contributing to the major absolute differences between samples. In this study, there were two ways to present the differences of six samples after the rlog transformation. The first mean is to calculate the Euclidean distance between each two samples in counting of all genes in the genome and visualize in the heatmap (Figure 4.4) and the second one is to cluster the top significantly down-regulated genes in the two strains based on their expression level and visualize them in a heatmap (Figure 4.5). The results from both means were consistent in clustering the three biological replicates of each strain together and different from the three biological replicates of the other strain.

To further validate the quality improvement of the 2017-version annotation, the same differential analysis workflow was performed but using the 2009-version annotation as the reference for the HTseq-count step. There were 819 genes significantly down-regulated with a log2 fold change less than -2 and 245 genes significantly upregulated with a log2 fold change larger than +2 while using the 2017-version annotation as reference (Table 4.3). On the contrary, 958 and 251 genes were found significantly down-regulated and upregulated in the same level while using the 2017-version annotation as reference (Table 4.3).

Significantly regulated KEGG metabolism pathways in Δ*cpvib-1* strain

As one of the GSA tools, GAGE was designed to reveal the relevant regulatory mechanisms from the transcriptome scale based differential expression analyses. In this study, the transcriptomes of Δ*cpvib-1* strain was the treated group and the EP155 strain was the untreated one with the expression level of each gene normalized by DESeq2

mentioned above. After the ID conversion using BLASTp and IDcoversion.py, 7,079 out of 11,171 genes returned the *N. crassa* KEGG ID compatible with the GAGE and Pathview tools. The normalized expression level of each gene with a recognizable KEGG ID matrix was fed to the GAGE with the *N. crassa* metabolism pathways KEGG dataset as reference.

In the end, there were 91 metabolic pathways summarized to be different with statistical data in the Δ*cpvib-1* strain compared to the EP155 stain. Twenty-nine of them were down-regulated and 62 up-regulated, respectively. Furthermore, GAGE provides a z-test to compare the net effect of differential expressed gene while using the as.group option rather than the paired_sample option leading to a relatively larger but more reliable p value [89]. There were two metabolic pathways highlighted as significantly down-regulated (p.value <0.1, highlighted in green) and 19 significantly up-regulated (p.value <0.1, highlighted in red) in the Δ*cpvib-1* strain from the GAGE analysis (Table 4.4).

In the down-regulated pathway group, the top four were the RNA transport, Ribosome, Aminoacyl-tRNA biosynthesis, and the mRNA surveillance pathways are all down-regulated with the relatively lower statistics means (stat.mean: mean of the gene-set transcription level changes) from -0.27 to -0.77 in the Δ*cpvib-1* strain compared to the EP155 strain (Table 4.4). The next four were the Proteasome, Arginine and proline metabolism, Ubiquitin mediated proteolysis and the Alanine, aspartate and glutamate metabolism pathways with the statistics means from -0.18 to -0.2 in the Δ*cpvib-1* strain while compared to the EP155 strain (Table 4.4).

In the up-regulated pathway group, four of the top five were the Biosynthesis of antibiotics, Biosynthesis of secondary metabolites, Carbon metabolism, and the Biosynthesis of amino acids pathways (Table 4.4). All of the listed pathways are the broad metabolism pathways composed of many individual pathways and highlighted here because the clustering of those individual pathways. Except for the above three, the most up-regulated pathway in the Δ*cpvib-1* strain is the Glycolysis / Gluconeogenesis pathway (Table 4.4). Furthermore, in all the 15 individual significantly up-regulated pathways, eight of them were the individual carbon metabolic pathways (Table 4.4).

In contrast, while using the 2009-verison annotation as reference in HT-seq-count and the same following procedures described above, in the ID conversion step, 7,007 out of 11,609 genes returned the *N. crassa* KEGG ID compatible with the GAGE. There were three metabolism pathways highlighted as significantly down-regulated (p.value <0.1, highlighted in green) and five significantly up-regulated (p.value <0.1, highlighted in red) in the Δ*cpvib-1* strain (Table 4.5). In the five significantly up-regulated pathways highlighted in red, there was only one individual carbon metabolism pathway (Table 4.5).

The Pathview package was used to visualize the pathways of interest in this study. First, in the up-regulated pathways group in the Δ*cpvib-1* strain, the ncr00010 Glycolysis / Gluconeogenesis pathway was presented with all critical enzymes in one string regulated consistently (Figure 4.6). The ncr00500 Starch and sucrose metabolism, the ncr00052 Galactose metabolism, and the ncr00680 Methane metabolism that are all linked to the glycolysis in the bigger view were presented with most critical enzymes consistently flowing to the D-glucose (Figure 4.7, Figure 4.8, Figure 4.9). One special pathway was presented here was the Oxidative phosphorylation metabolism (Figure

4.10). Besides, the ncr03013 RNA Transport pathway and ncr04120 Ubiquitin mediated proteolysis pathway was presented as the examples for the down-regulated pathways group in the Δ*cpvib-1* strain (Figure 4.11, Figure 4.12).

Significantly expressed genes' GO enrichment in Δ*cpvib-1* strain

REVIGO was used to cluster the GO terms based on the functional similarity by feeding with a matrix of the GO term and its associated log2 fold change. The table showed the most frequent GO terms in the significantly regulated genes in the Δ*cpvib-1* strain (Table 4.6). The results showed a small number of top-ranking GO terms with a dispensability value as 0 were the GO:0003700 transcription factor activity, GO:0003884 D-amino-acid oxidase activity, GO:0005375 copper ion transmembrane transporter activity and the GO:0045735 nutrient reservoir activity and all having the log2 fold change value less than -2, except the GO:0005096 GTPase activator activity with a log2 fold change value as 3.23 (Table 4.6). Moreover, there were some more GO terms related to carbon and lipid metabolism and DNA polymerase activities that were listed in the table as well drew the attention because of the phenotype shift in the Δ*cpvib-1* strain and former studies of VIB-1 in *N. crassa* (Table 4.6).

In addition to a table format, REVIGO provides three additional visualization strategies. First, the scatterplot was draw based on the semantic similarities of the GO terms (x axis) and the log2 fold change values (y axis). The scatterplot here showed the up-regulated clusters (Red color bubbles) were the Oligopeptide transport, Carbohydrate metabolism, Proteolysis, Carboxylic acid metabolism, etc. (Figure 4.13). Meanwhile, the down-regulated clusters (Blue-green bubbles) were the Pathogenesis, D-amino acid

metabolism, lipid catabolism, RNA processing, iron transmembrane transport, and others (Figure 4.13).

The second presentation method is the graph-based clusters interaction view, which using a node to represent a cluster from the scatterplot and use the edges to represent at least 3% of the strongest GO term pairwise similarities. The interactive graph showed the level of the transcription regulating was related to RNA process and the nucleoside metabolism, which related to the carbon, lipid and protein metabolism (Figure 4.14). On the other side, another interactive net was listed here of different molecule transport processes, which were highlighted and linked with each other (Figure 4.15).

The last presentation method is the treemap showing the two-level hierarchy of GO terms by joining the clusters from the scatterplot and interactive graph to several high-level groups. In the treemap graph, each rectangle represents a single cluster, which were joined into the superclusters in different colors. It was obvious that the D-amino acid metabolism was the most major supercluster regulated in the Δ*cpvib-1* strain (Figure 4.20). The response to stress, pathogenesis, molecular transport, and carbon metabolism were highlighted as the relatively smaller superclusters, which were consistent with the phenotype observed in the absence of *cpvib-1* (Figure 4.16).

Discussion

RNA-Seq is now widely used to reveal the transcriptome profile in organisms and can be essential for interpreting the functional elements of the genome and revealing the molecular constituents of the metabolic pathways. The depth of data assisted the re-annotation of the genome by providing accuracy and improved gene model predictions.

Here, it was utilized to provide insights into the regulation mechanisms of the transcriptional activator CPVIB-1 in the plant pathogen *C. parasitica.*

As mentioned in Chapter II concerning the quality of the RNA-Seq reads and their alignment percentages to the genome, both the transcriptome of the Δ*cpvib-1* and EP155 strains shared similar high-quality results as described in Results of this Chapter. Furthermore, with the advanced rlog strategy to remove the bias of the low count reads from the biological replicates of two strains, the sample transcriptome profiles' similarity was displayed in the Euclidean distance heatmap and expression level heatmap, indicating their consistency within the same strain and differences among the two strains. The above results provided evidence to confirm the accuracy of the RNA-Seq platform and credibility of the library preparation of the samples.

CPVIB-1 was found to be a universal transcriptional activator itself by the differential expression analysis using DESeq2 after the normalization and statistics analysis. There were 1,064 genes (approximately 10% of the genome) found significantly regulated with a p-value < 0.05 and log2 fold change's absolute value > 2 in the Δ*cpvib-1* strain (Table 4.3). Moreover, the GO term clusters summarized from the significantly regulated genes in the Δ*cpvib-1* strain (Table 4.6) showed great diversity. One of the highlighted GO term is GO:0003700 transcription factor activity with a mean log2 fold change value as -2.3627 indicating the transcript level of various transcription factors were down-regulated in the absence of the *cpvib-1* gene in *C. parasitica* (Table 4.6). This result was found to be consistent with the phenotype shift of the Δ*cpvib-1* strain resulting in the slower hyphal growth rate, profuse conidiation and increased pigmentation caused by broad cell cytoplasm processes.

CPVIB-1 was discovered to mediate the carbohydrate catabolite metabolism and other nutrient catabolite metabolism by affecting the hydrolytic-activity-related genes. Two sets of results were found to support this. First, there were five various but directly related specific KEGG carbon pathways found significantly up-regulated in the Δ*cpvib-1* strain (Table 4.4). The starch and sucrose metabolism taking both the polysaccharides and sucrose as the sugar resource to produce the D-glucose (Figure 4.7). Also, as the product of the methane metabolism as well as the fructose and mannose metabolism, the fructose 6-P is convertible with D-glucose (Figure 4.9). Moreover, both the pentose phosphate pathway and galactose metabolism are also leading to the product of D-glucose, which is also reversible (Figure 4.8). Almost all enzymes that are comprising the glycolysis/gluconeogenesis pathway were regulated in the Δ*cpvib-1* strain and they were consistently regulated in the same metabolic flow direction to consume the glucose to the citrate cycle for ATP synthesis (Figure 4.6).

During the conversion of KEGG ID's from *C. parasitica* to *N. crassa*, we found that only 7096 genes were mapped to the KEGG pathways, presumably due to *N. crassa* possessing different pathways. Given *N. crassa* is not a pathogen, pathways of specific interest to the understanding of such behavior could be missed and thus they would not be highlighted. Additionally, therefore, we use REVIGO term enrichment analysis to provide additional global information. In the Δ*cpvib-1* strain transcriptome, the results from the clustered GO terms in scatterplot indicated the up-regulation of carbohydrate, carboxylic acid and amine catabolism metabolism as well as the down-regulation of lipid and D-amino acid metabolism with the log2 fold change values larger than +4 (Figure 4.13). Meanwhile, the super-group treemap graph of the clustered GO terms in the

Δ*cpvib-1* strain transcriptome, the major super-group that were significantly regulated were the nutrient metabolism, molecular transport (Figure 4.16). In *N. crassa,* VIB-1 was found to repress the glucose signaling and encourage the carbon catabolite repression process, thus enabling a proper cellular response for plant biomass deconstruction [34]. Therefore, it was not surprising that the Δ*cpvib-1* strain resulted in encouraging the utilization of carbon catabolite and glucose signaling process without the presence of the CPVIB-1.

CPVIB-1 was also found to be a mediator to the pathogenesis of *C. parasitica.* Although the KEGG pathway analysis did not highlight pathways in the Δ*cpvib-1* strain related to the pathogenesis, the GO term enrichment analysis highlighted a cluster of genes related to the pathogenesis with a mean log2 fold change value as -6 (Figure 4.13). Additionally, the super-group treemap graph of the clustered GO terms in the Δ*cpvib-1* strain transcriptome highlighted the pathogenesis super-group (in pink color) as one of the major significantly regulated super-groups (Figure 4.16). In conclusion, the down-regulation of pathogenesis related genes in the Δ*cpvib-1* strain was consistent with the virulence attenuation phenotype we found in the virulence assay of the Δ*cpvib-1* strain.

CPVIB-1 was discovered to be essential for proper RNA processing and biosynthesis. Both the KEGG pathway analysis and GO term enrichment analysis were consistent in presenting the significant down-regulation these pathways in the Δ*cpvib-1* strain. Similar results were reported in *N. crassa* that VIB-1 mutant strain displayed significantly reduced hyphal growth rate and reduced cellulose consumption level.

The absence of CPVIB-1 triggered the higher transcription level of genes responding to oxidative stress, the oxidation reduction process, and oxygen binding. In

the transcriptome of the Δ*cpvib-1* strain, the oxidative phosphorylation pathway, oxidation reduction process and oxygen binding GO term cluster were both found significantly up-regulated. Without CPVIB-1 the balance of producing and eliminating reactive oxygen species (ROS) was disturbed leading to the oxidative stress which can lead to the severe damage of DNA, protein and lipids [94]. In response to the stress in the absence of CPVIB-1, the genes related to response to the oxidative stress were transcribed in higher level.

With all the above results, it is now clear that the absence of the key universal transcription activator CPVIB-1 caused the imbalance of stress response systems, altered DNA, protein and lipids synthesis, disturbed various transcription factor activities and also RNA processing. The attenuated pathogenesis feature might be caused by the glucose signaling pathway shift in the Δ*cpvib-1* strain preventing the fungus from utilizing the plant cell wall components, like cellulose. These results indicate that CPVIB-1 is a much more global regulator than was first suspected, and is critical for the proper function of many cellular activities that may be unrelated to its role in vegetative incompatibility.

Figures and Tables

Table 4.1 RNA-Seq read quality of three biological replicates from the *C. parasitica Δcpvib-1*strain.

Sample	Reads Yield (Million)	*% of >= Q30 Bases	Mean Quality Score
Δ*cpvib-1*s1	47.00	94.74	35.75
Δ*cpvib-1*s2	24.27	94.12	35.63
Δ*cpvib-1*s3	33.99	93.82	35.57

*% of >= Q30 Bases means the percentage of all bases in a read containing no more than one error in each 1,000 bases [68; 75].

Table 4.2 The quality of read alignments from Δ*cpvib-1* strain.

Sample	Reads Yield (Million)	Reads number mapped to genome	[a]Concordant pair alignment rate	[b]Discordant alignments rate
Δ*cpvib-1*s1	47.00	44.86 M (95.45%)	92.5%	0.4%
Δ*cpvib-1*s2	24.27	23.53 M (96.95%)	93.5%	0.6%
Δ*cpvib-1*s3	33.99	32.98 M (97.02%)	93.4%	0.7%

[a] Concordant pair alignment rate, referring to the percentage of reads from both paired-end sequencing file matched to the same locus in the genome.
[b] Discordant alignment rate, referring to the percentage of reads from both paired-end sequencing file did not match to the same locus in the genome.

Table 4.3 The number of significantly regulated genes in the Δ*cpvib-1* strain compared to EP155 strain.

Gene number	2017-version	2009-version
Up-regulated	245	251
Down-regulated	819	958
Nonzero reads counts	10,755	11,209

These are significantly regulated gene number with a p adjusted value less than 0.05 and a log2 fold change larger than 2 or less than -2.

Table 4.4 The metabolic pathways regulated in Δcpvib-1strain compared to EP155 strain with 2017-version annotation as reference.

Down-regulated pathway name	stat.mean	p.val	Up-regulated pathway name	stat.mean	p.val
ncr03013 RNA transport	-0.77	0.09	ncr01130 Biosynthesis of antibiotics	2.80	0.00
ncr03010 Ribosome	-0.73	0.10	ncr01110 Biosynthesis of secondary metabolites	2.64	0.00
ncr00970 Aminoacyl-tRNA biosynthesis	-0.31	0.30	ncr00010 Glycolysis / Gluconeogenesis	2.78	0.00
ncr03015 mRNA surveillance pathway	-0.27	0.32	ncr01200 Carbon metabolism	2.37	0.00
ncr03050 Proteasome	-0.20	0.37	ncr01230 Biosynthesis of amino acids	2.21	0.00
ncr00330 Arginine and proline metabolism	-0.18	0.38	ncr00500 Starch and sucrose metabolism	2.06	0.00
ncr04120 Ubiquitin mediated proteolysis	-0.18	0.38	ncr00052 Galactose metabolism	1.80	0.00
ncr00250 Alanine, aspartate and glutamate metabolism	-0.18	0.38	ncr00680 Methane metabolism	1.79	0.00
ncr04113 Meiosis - yeast	-0.15	0.40	ncr00030 Pentose phosphate pathway	1.70	0.00
ncr00340 Histidine metabolism	-0.14	0.41	ncr00051 Fructose and mannose metabolism	1.53	0.00
ncr04111 Cell cycle - yeast	-0.13	0.41	ncr00520 Amino sugar and nucleotide sugar metabolism	0.98	0.05
ncr03022 Basal transcription factors	-0.13	0.41	ncr00260 Glycine, serine and threonine metabolism	0.98	0.05
ncr00410 beta-Alanine metabolism	-0.13	0.41	ncr00190 Oxidative phosphorylation	0.92	0.06
ncr00240 Pyrimidine metabolism	-0.12	0.42	ncr00040 Pentose and glucuronate interconversions	0.93	0.06
ncr00760 Nicotinate and nicotinamide metabolism	-0.11	0.43	ncr00350 Tyrosine metabolism	0.85	0.07
ncr00220 Arginine biosynthesis	-0.09	0.44	ncr00071 Fatty acid degradation	0.83	0.08
ncr03020 RNA polymerase	-0.09	0.44	ncr00230 Purine metabolism	0.78	0.09
ncr00380 Tryptophan metabolism	-0.08	0.45	ncr03018 RNA degradation	0.75	0.10
ncr04136 Autophagy - other	-0.07	0.45	ncr00620 Pyruvate metabolism	0.63	0.14
ncr00310 Lysine degradation	-0.07	0.45	ncr01040 Biosynthesis of unsaturated fatty acids	0.63	0.14
ncr00280 Valine, leucine and isoleucine degradation	-0.07	0.45	ncr01212 Fatty acid metabolism	0.55	0.17
ncr04011 MAPK signaling pathway - yeast	-0.07	0.46	ncr00100 Steroid biosynthesis	0.50	0.20
ncr04138 Autophagy - yeast	-0.05	0.47	ncr00562 Inositol phosphate metabolism	0.48	0.20

Table 4.5 The metabolism pathways regulated in the Δ*cpvib-1*strain compared to EP155 strain with 2009-version annotation as reference.

Down-regulated pathway name	stat.mean	p.val	Up-regulated pathway name	stat.mean	p.val
ncr03010 Ribosome	-1.89	0.00	ncr01100 Metabolic pathways	1.29	0.01
ncr03030 DNA replication	-1.11	0.03	ncr00010 Glycolysis / Gluconeogenesis	0.97	0.05
ncr03430 Mismatch repair	-0.80	0.09	ncr01210 2-Oxocarboxylic acid metabolism	0.89	0.07
ncr03420 Nucleotide excision repair	-0.65	0.14	ncr03040 Spliceosome	0.85	0.07
ncr00240 Pyrimidine metabolism	-0.60	0.15	ncr01110 Biosynthesis of secondary metabolites	0.84	0.07
ncr04146 Peroxisome	-0.32	0.29	ncr01130 Biosynthesis of antibiotics	0.77	0.09
ncr04145 Phagosome	-0.16	0.39	ncr03013 RNA transport	0.73	0.11

Table 4.6 The top rank GO term summarized in the REVIGO from the significantly regulated genes in the *Δcpvib-1* strain.

term_ID	description	frequency[c]	value[a]	uniqueness[b]	dispensability
GO:0003700	transcription factor activity, sequence-specific DNA binding	4.22%	-2.3627	0.974	0
GO:0003884	D-amino-acid oxidase activity	0.02%	-10.9393	0.909	0
GO:0005096	GTPase activator activity	0.18%	3.2366	0.985	0
GO:0005375	copper ion transmembrane transporter activity	0.03%	-5.0129	0.948	0
GO:0005488	binding	55.66%	3.0819	0.993	0
GO:0010181	FMN binding	0.70%	-7.5603	0.893	0
GO:0045735	nutrient reservoir activity	0.04%	-4.8423	0.985	0
GO:0051907	S-(hydroxymethyl)glutathione synthase activity	0.00%	-4.9121	0.926	0.015
GO:0003963	RNA-3'-phosphate cyclase activity	0.01%	-4.1529	0.936	0.016
GO:0016742	hydroxymethyl-, formyl- and related transferase activity	0.23%	-8.6942	0.884	0.02
GO:0004476	mannose-6-phosphate isomerase activity	0.03%	3.1358	0.949	0.02
GO:0004806	triglyceride lipase activity	0.04%	-8.109	0.836	0.021
GO:0016874	ligase activity	3.54%	-3.1298	0.942	0.03
GO:0016787	hydrolase activity	22.29%	2.1006	0.935	0.093
GO:0003743	translation initiation factor activity	0.41%	2.9937	0.947	0.1
GO:0003887	DNA-directed DNA polymerase activity	0.46%	-2.5228	0.863	0.24

[a] Value column is representing the mean log2 fold change values for each GO term.
[b] Uniqueness column is representing the similarity measurement of the GO term in the cluster.
More unique term tends to be less dispensable.
[c] Frequency is the percentage of humans proteins in UniProt were annotated with a GO term in the GOA database, i.e., a higher frequency denotes a more general GO term, a lower frequency denotes a more specific GO term [92].

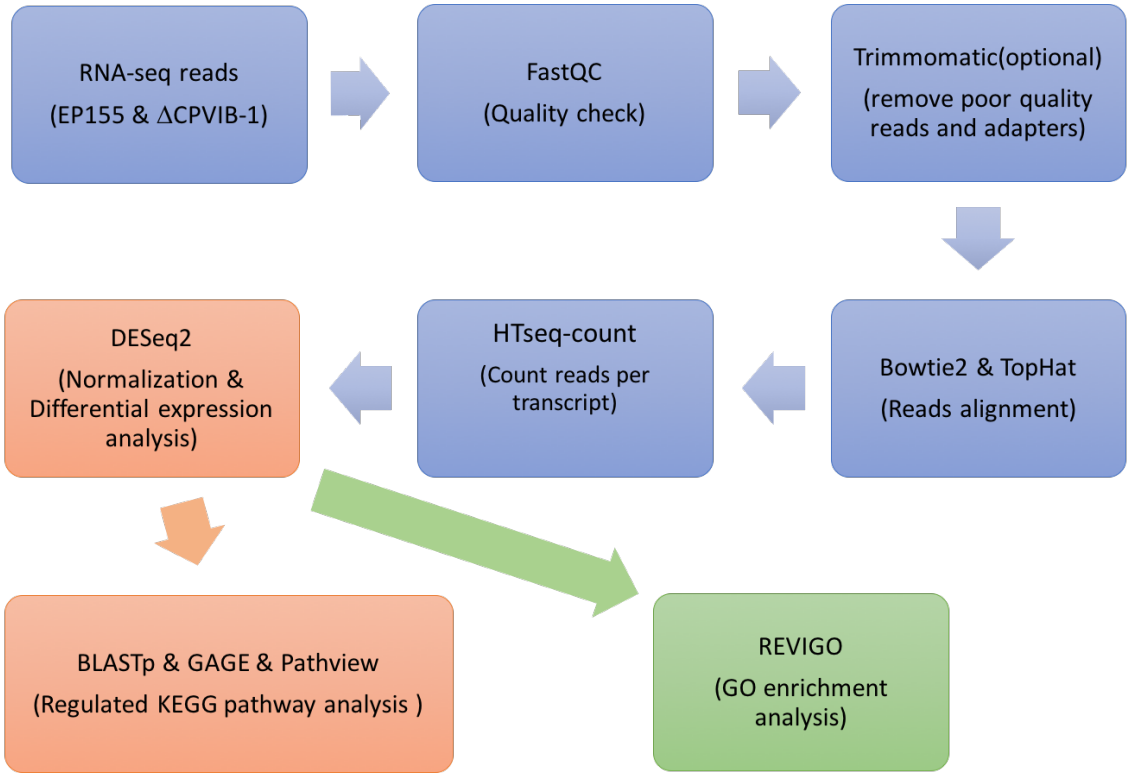


Figure 4.1 The differential expression analysis workflow.

Blue frames were included in the python on-step pipeline. Orange frames were included in the R script. Green frame was a web-server performance.

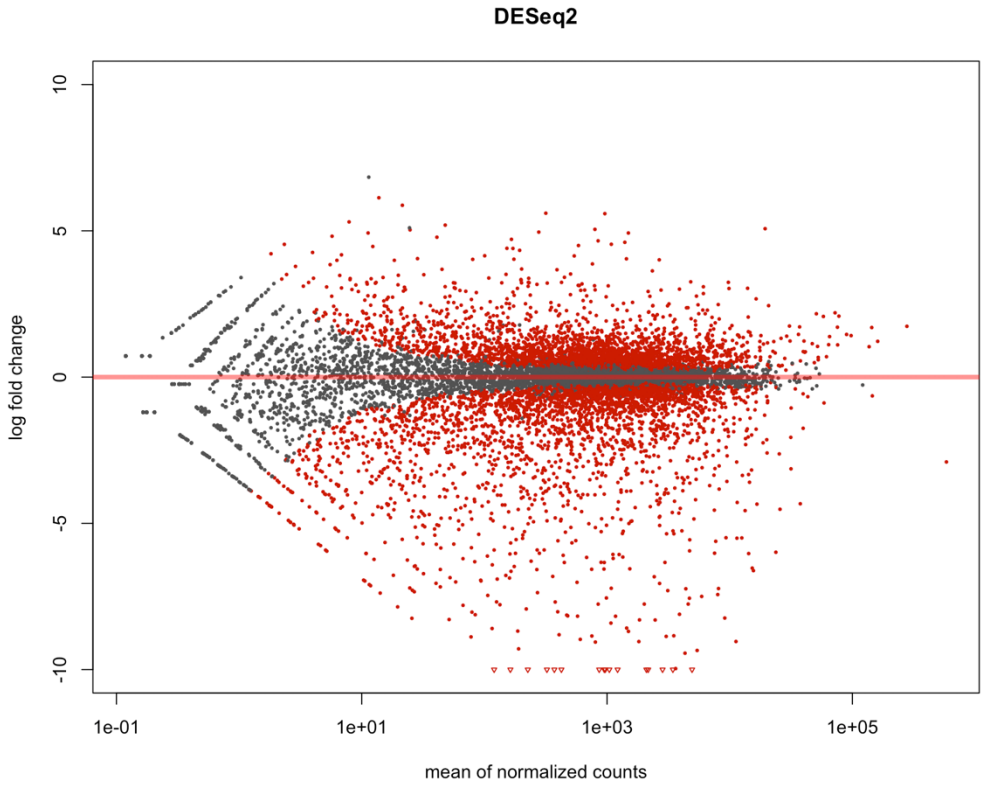


Figure 4.2 Log2 fold change plot of the Δ*cpvib-1* strain over the mean of normalized counts.

The plot represents each gene with a dot. The x axis is the average expression over all samples, the y axis is the log2 fold change between treatment and control. Genes with an adjusted p value below a threshold (here 0.1, the default) are shown in red. The red labeled triangles at the bottom represents the genes with a log2 fold change less than -10 and an adjusted p value below 0.1. This plot demonstrates that only genes with a large average normalized count contain sufficient information to yield a significant call.

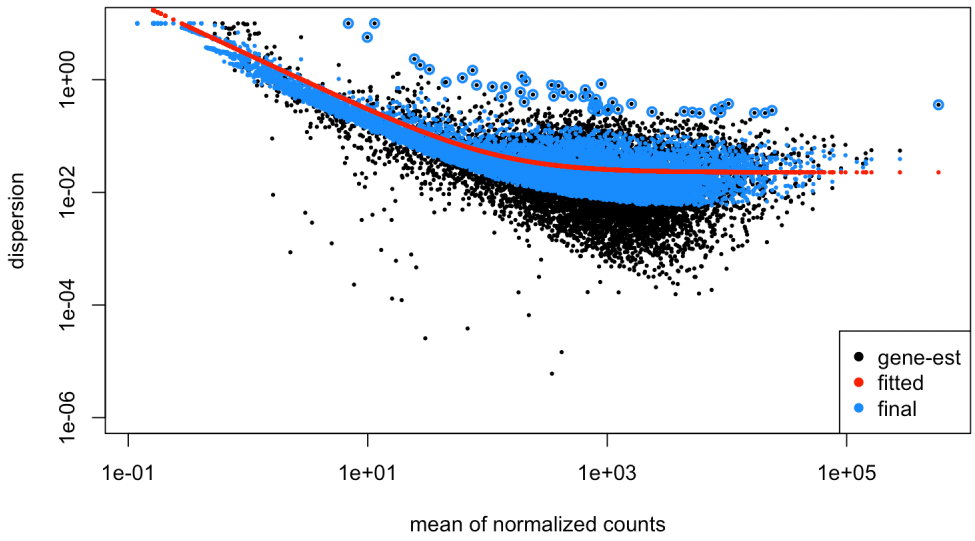


Figure 4.3 Dispersion plot of the Δ*cpvib-1* strain over the mean of normalized counts.

One black dot represents a gene's original dispersion and the mean of normalized counts. One blue dot represents the gene shrunken version of the dispersion and the mean of normalized counts. The red line is the trend of the mean of the dispersions of all genes. The blue circle represents the gene with a large dispersion that is not included in the shrunken process.

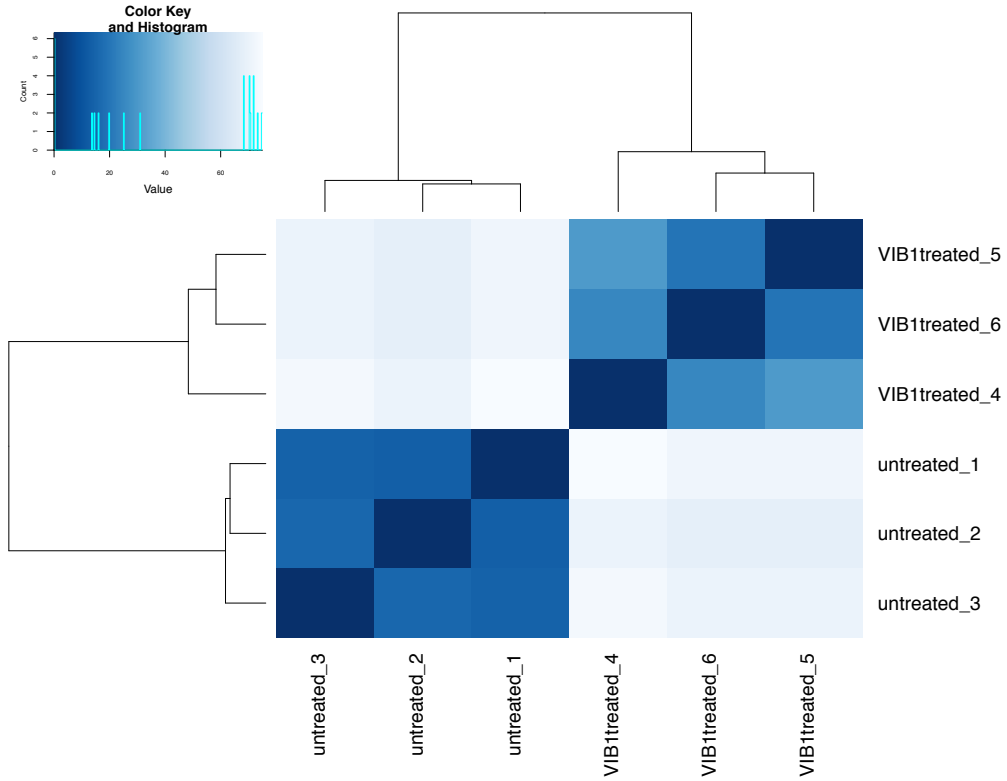


Figure 4.4 Heatmap of Euclidean sample distances of six samples in two strains after rlog transformation.

Three biological replicates for EP155 strain were labeled with untreated_1, untreated_2, untreated_3.
Three biological replicates for Δ*cpvib-1*strain were labeled with VIB1treated_1, VIB1treated_2, VIB1treated_3.

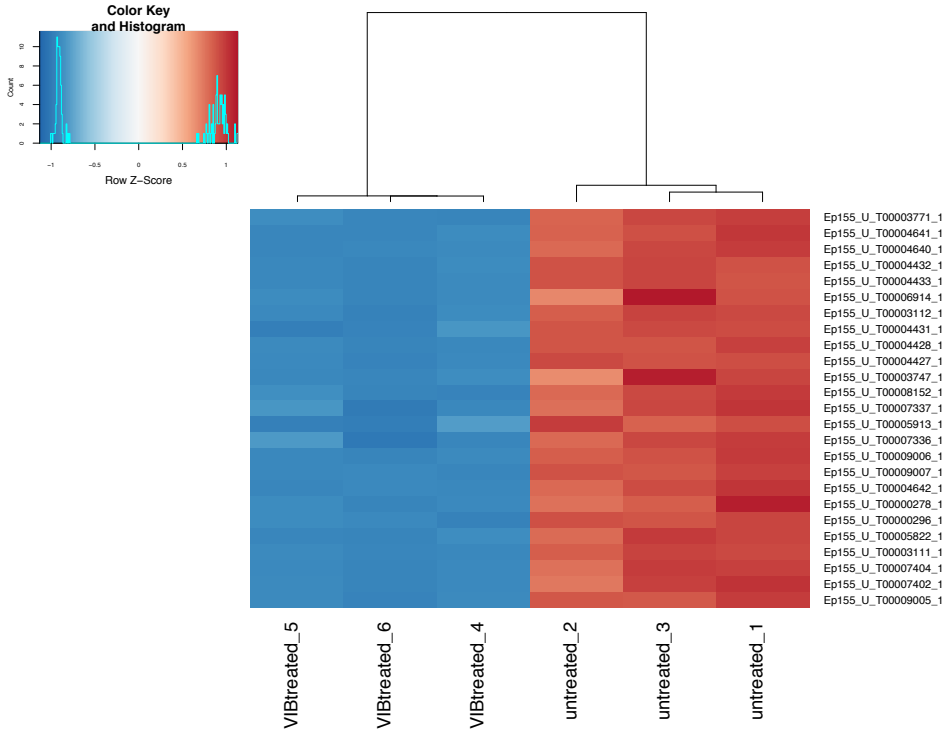


Figure 4.5 Heatmap of the top 25 significantly down-regulated genes clustering in six samples.

Three biological replicates for EP155 strain were labeled with untreated_1, untreated_2, untreated_3.
Three biological replicates for Δ*cpvib-1*strain were labeled with VIB1treated_1, VIB1treated_2, VIB1treated_3.

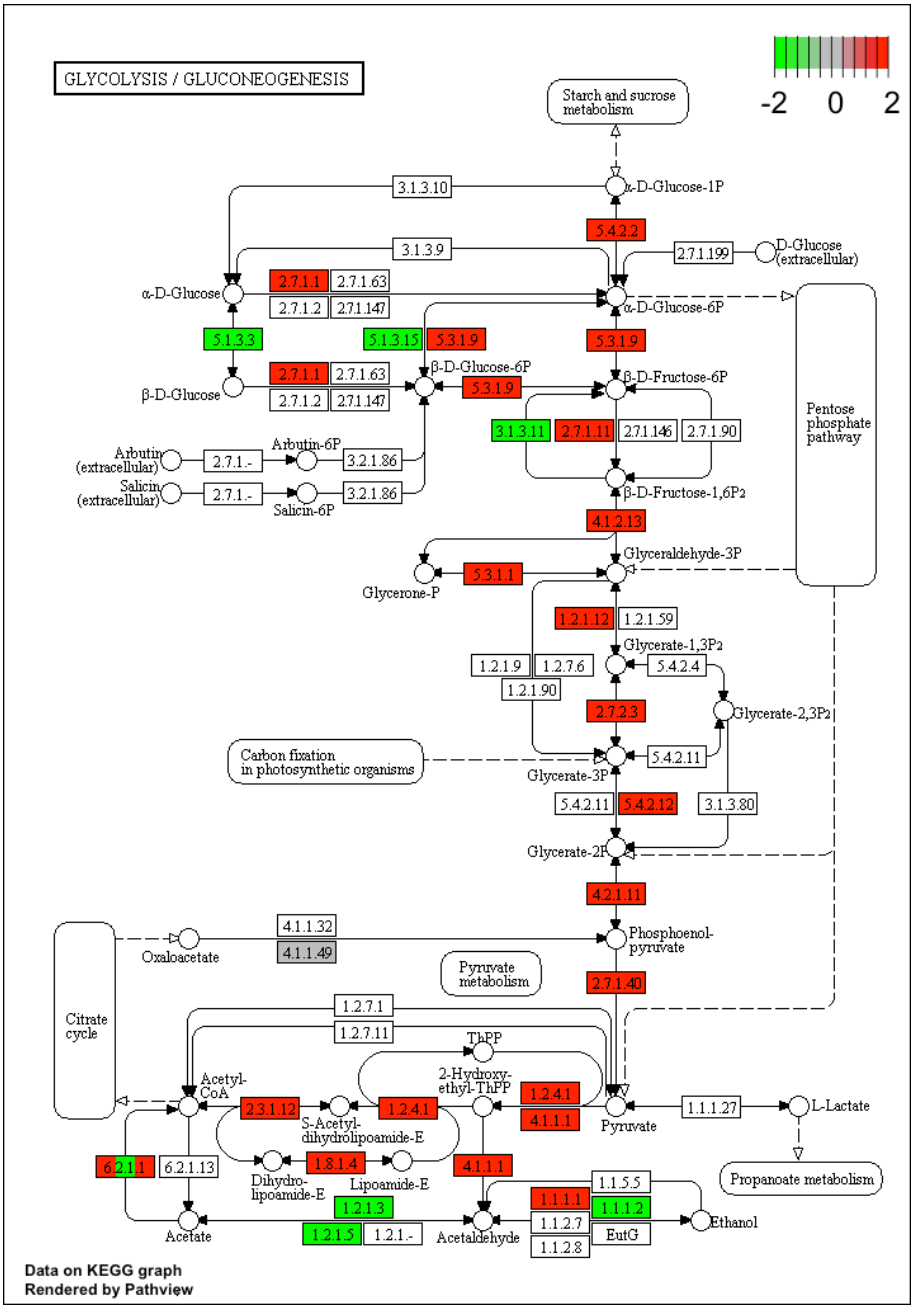


Figure 4.6 KEGG view of ncr00010 Glycolysis / Gluconeogenesis pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain.
Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain.
Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain.
Each white labeled frame stands for an enzyme not existing in *N. crassa*.

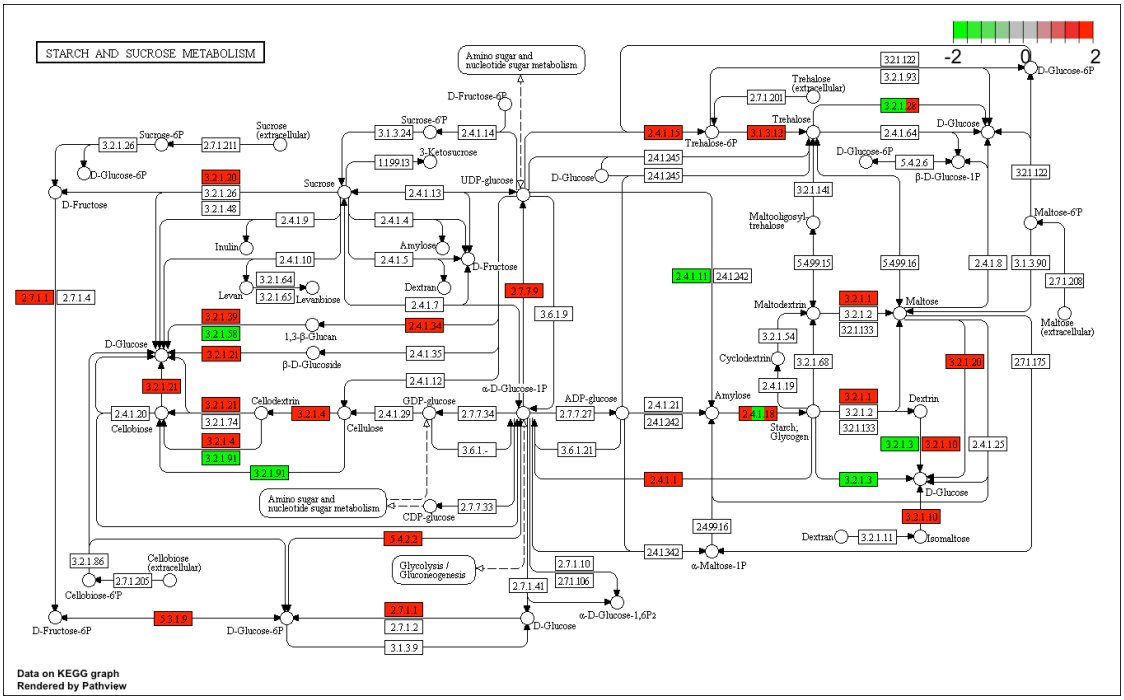


Figure 4.7 KEGG view of ncr00500 Starch and sucrose metabolism pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain.
Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain.
Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain.
Each white labeled frame strands for an enzyme not existing in *N. crassa*.

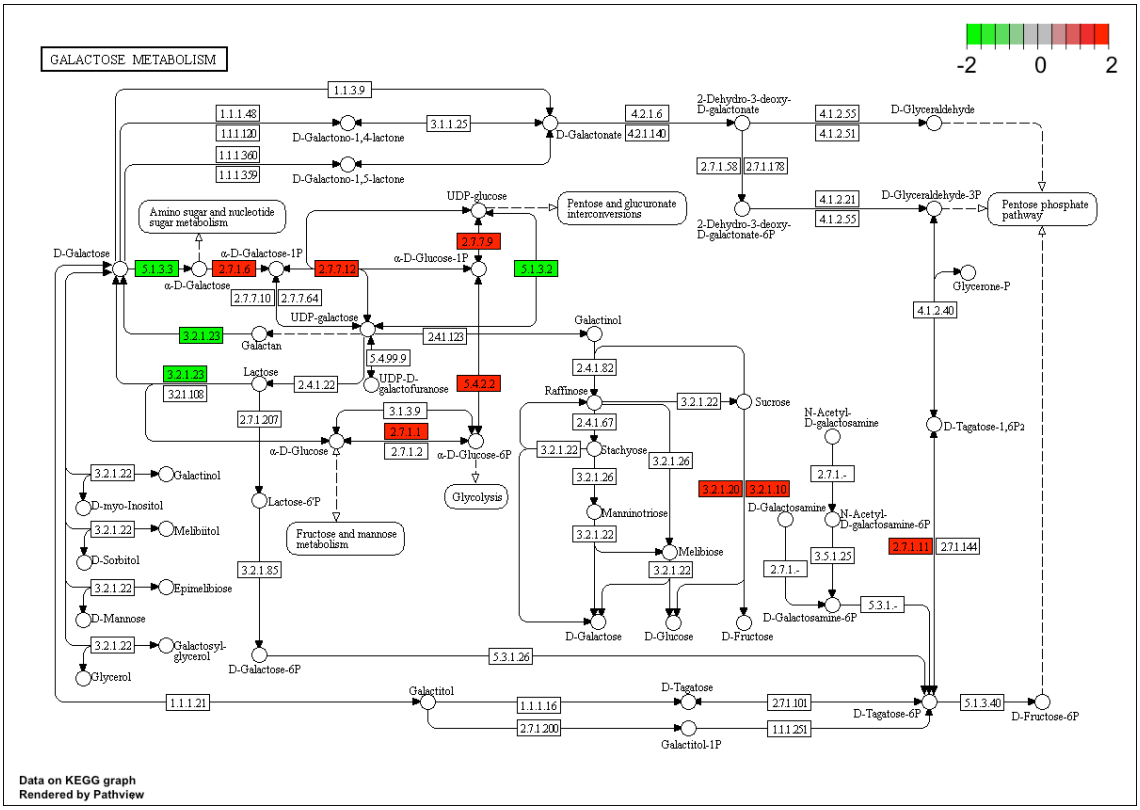


Figure 4.8 KEGG view of ncr00052 Galactose metabolism pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain.
Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain.
Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain.
Each white labeled frame strands for an enzyme not existing in *N. crassa*.

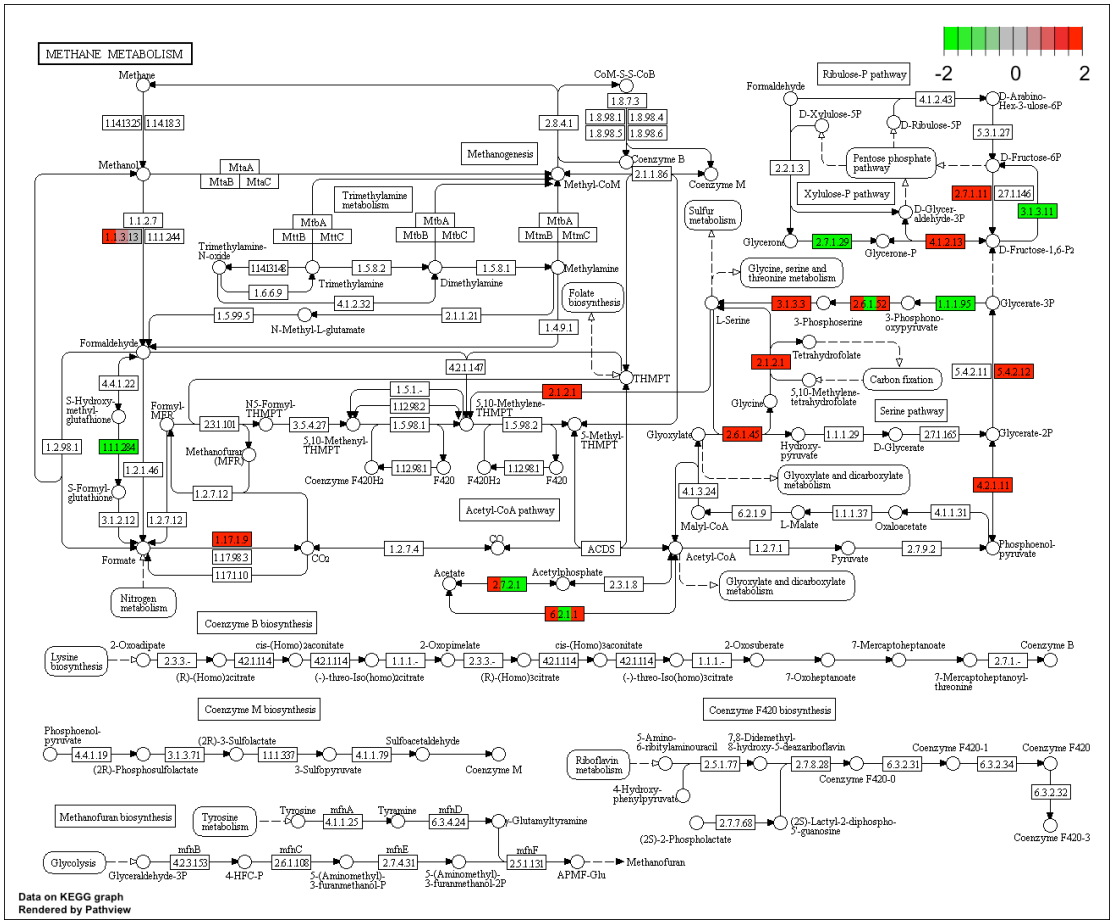


Figure 4.9 KEGG view of ncr00680 Methane metabolism pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain.
Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain.
Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain.
Each white labeled frame stands for an enzyme not existing in *N. crassa*.

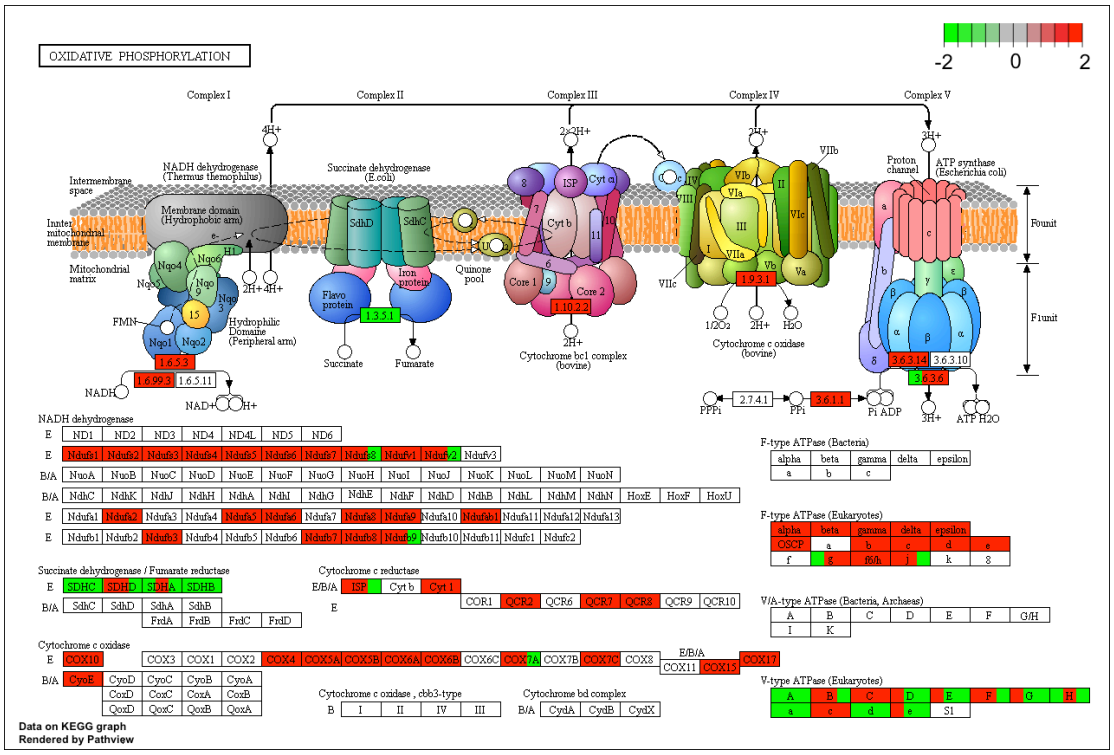


Figure 4.10 KEGG view of ncr00190 Oxidative metabolism pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain.
Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain.
Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain.
Each white labeled frame stands for an enzyme not existing in *N. crassa*.

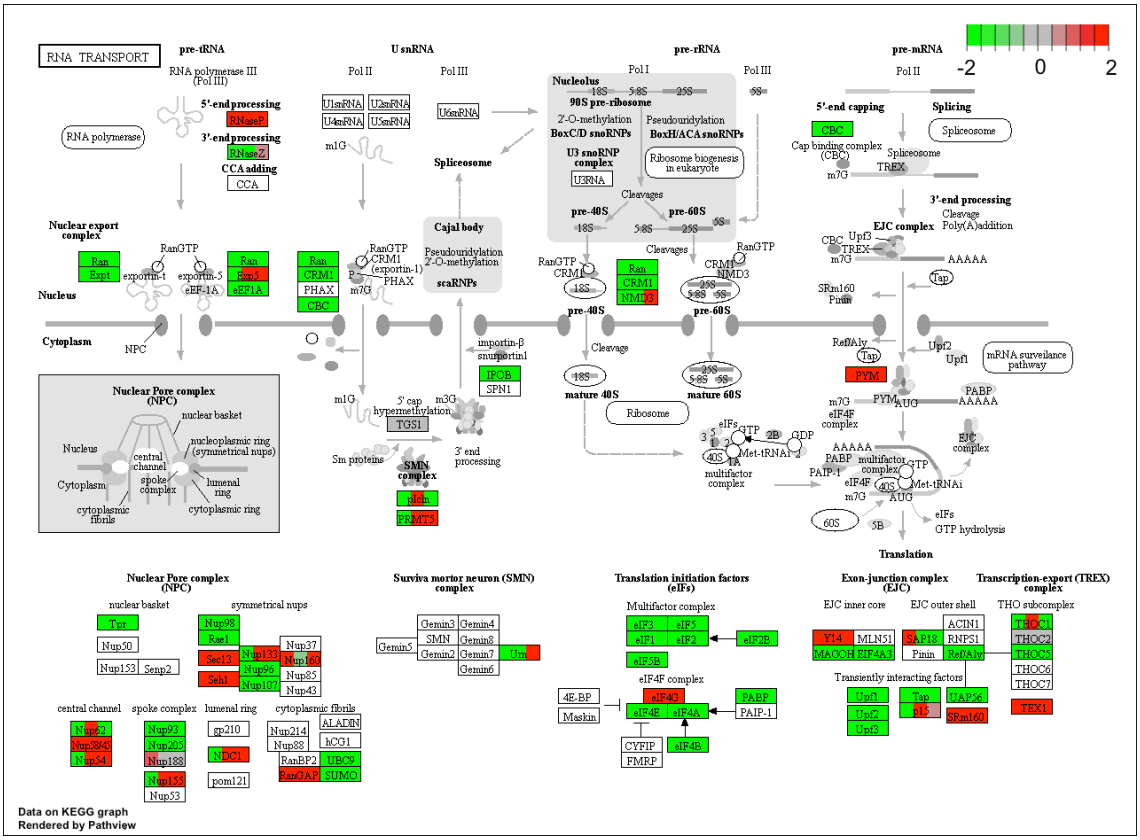


Figure 4.11 KEGG view of ncr03013 RNA Transport metabolism pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain. Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain. Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain. Each white labeled frame stands for an enzyme not existing in *N. crassa*.

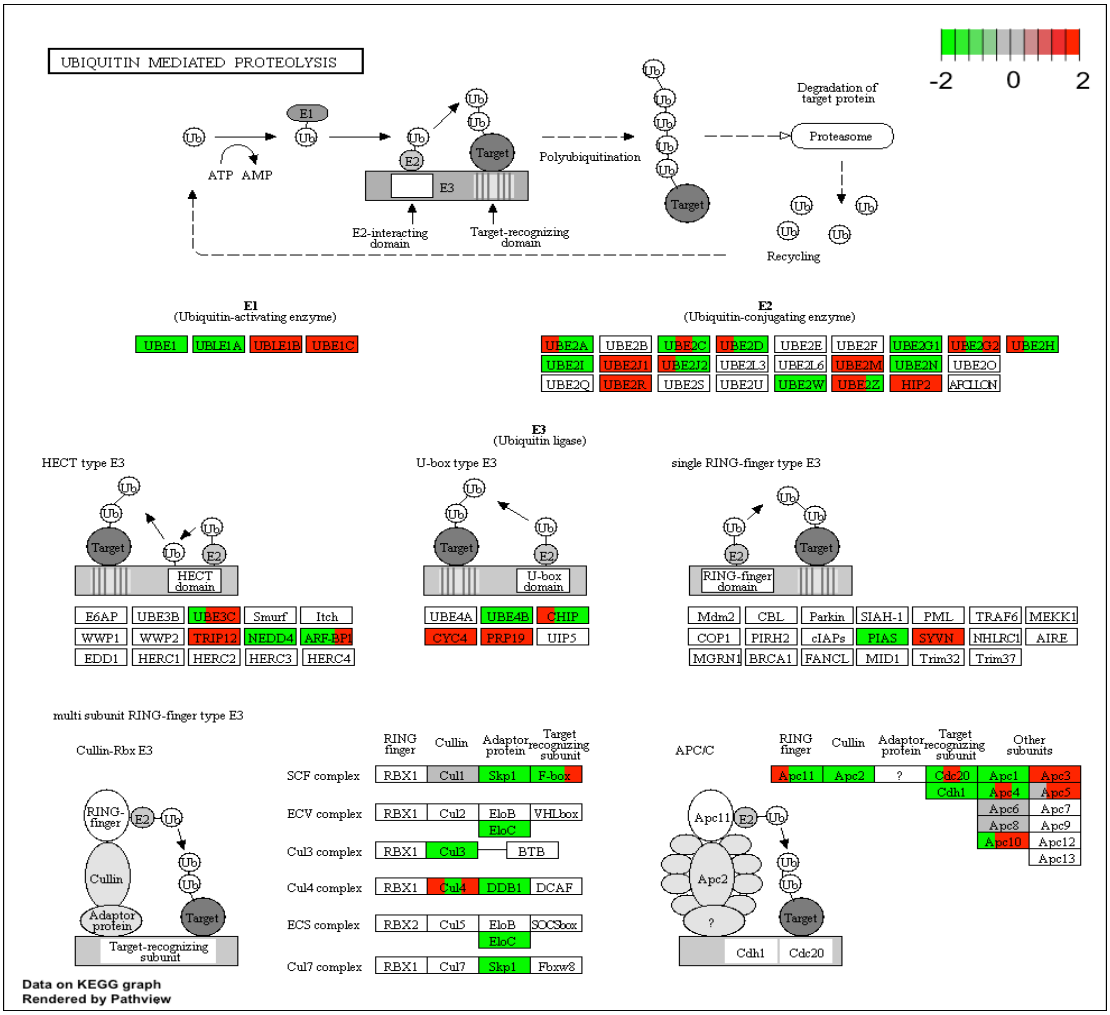


Figure 4.12 KEGG view of ncr04120 Ubiquitin mediated proteolysis pathway.

Each red labeled frame stands for an enzyme with a positive fold change in the Δ*cpvib-1* strain. Each green labeled frame stands for an enzyme with a negative fold change in the Δ*cpvib-1* strain. Each gray labeled frame stands for an enzyme with no fold change in the Δ*cpvib-1* strain. Each white labeled frame stands for an enzyme not existing in *N. crassa*.

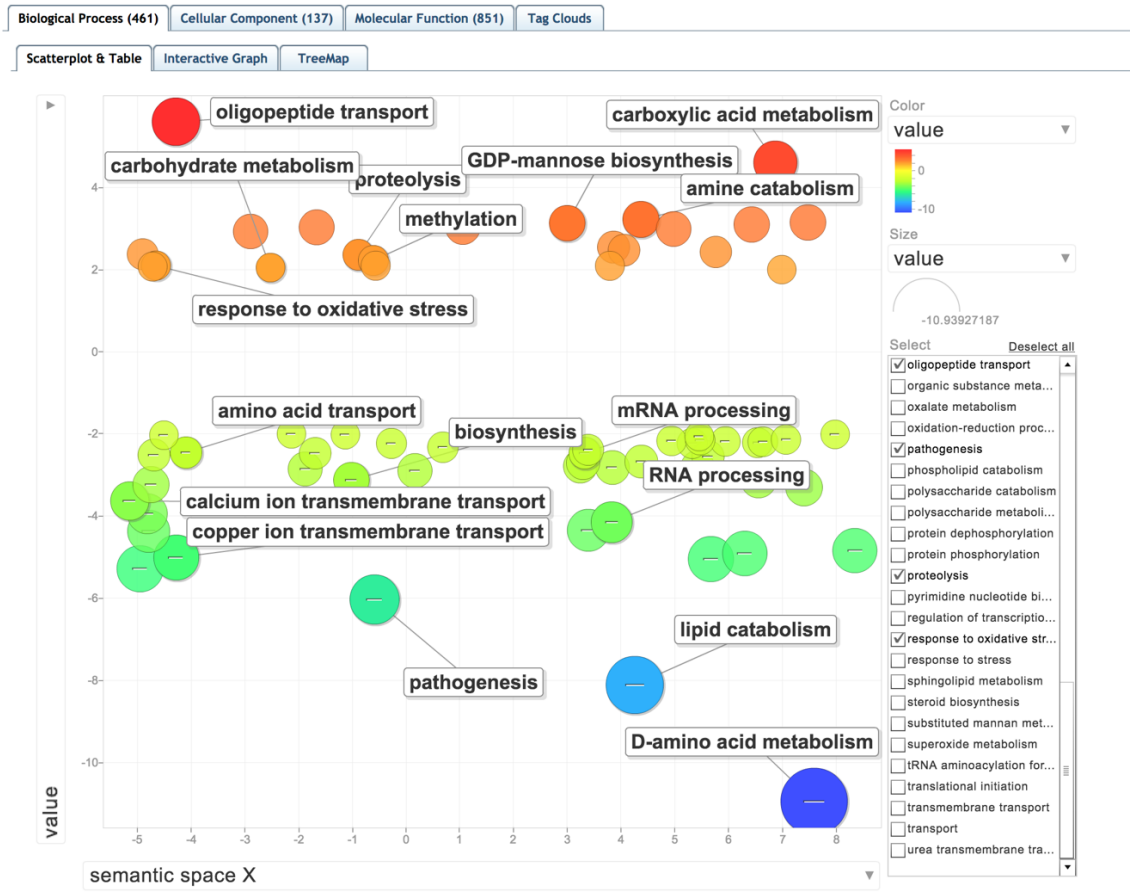


Figure 4.13 The "Scatterplot & Table" view of REVIGO showing the GO clusters of the significantly regulated genes in the Δ*cpvib-1* strain.

The y axis represents the log2 fold change value of each GO term. The x axis represents the semantic scale of the GO terms. The bubble color indicates the log2 fold change value from the positive (red, up-regulated) to negative (blue, green, and yellow down-regulated). The size of the bubble indicates the frequency of the GO term in the dataset.

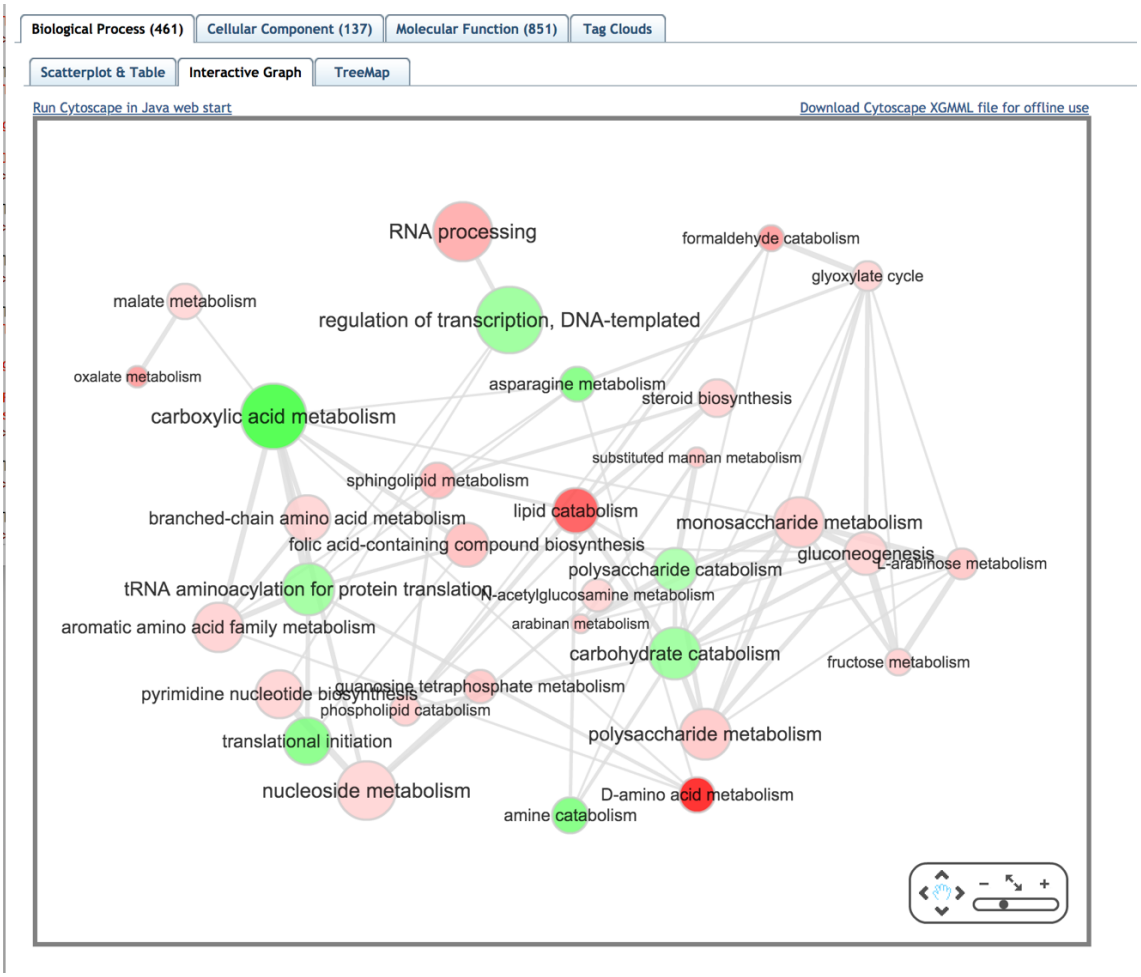


Figure 4.14 The “Interactive graph” view of REVIGO presenting the carbon, nitrogen metabolism clusters and their interactive RNA processing clusters in nodes and their interactive relationship by edges from the significantly regulated genes in the Δ*cpvib-1* strain.

The bubble size refers to the number of genes in the biological processes. The log2 fold change values were shown using color shading with negative values in red and positive values in green.

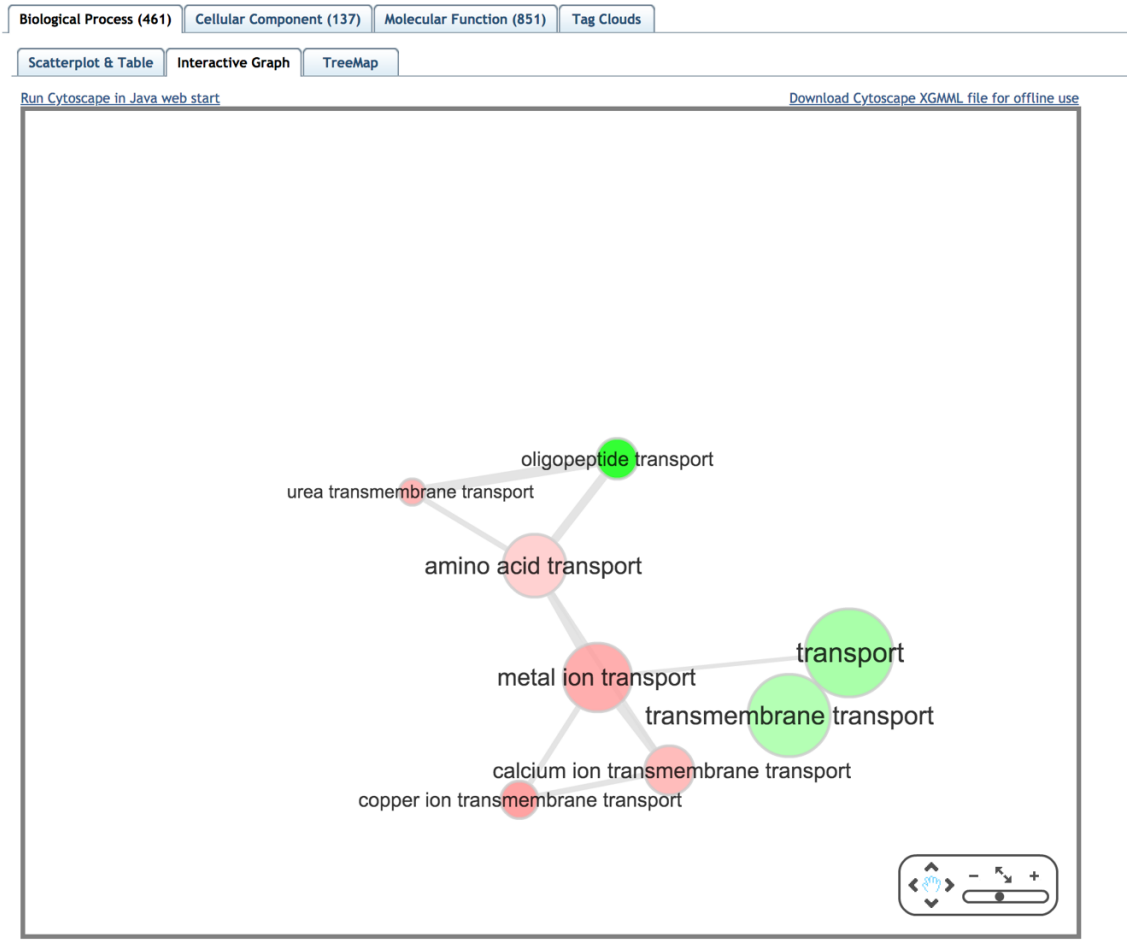


Figure 4.15 The "Interactive graph" view of REVIGO presenting the clusters of different transport process in nodes and their interactive relationship by edges from the significantly regulated genes in the Δ*cpvib-1* strain.

The bubble size refers to the number of genes in the biological processes. The log2 fold change values were shown using color shading with negative values in red and positive values in green.

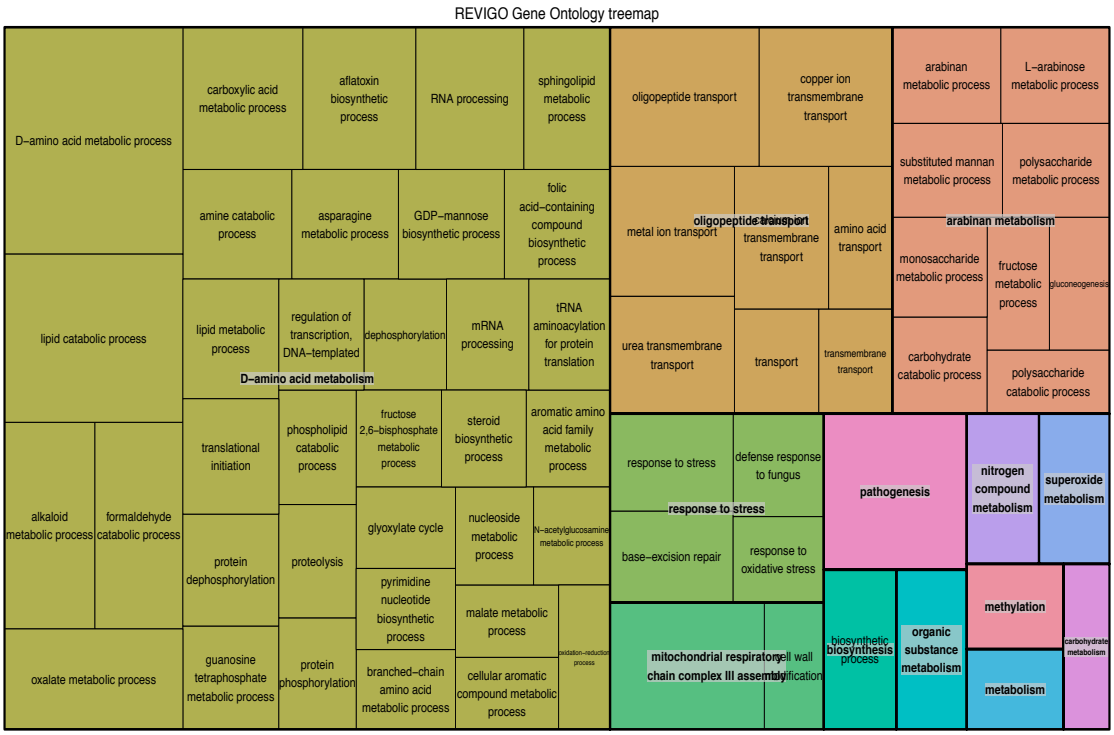


Figure 4.16 The “Treemap graph” view of REVIGO representing the supercluster groups in different colors.

One rectangle is representing a cluster from the scatterplot and interactive graph. Each color represents a supercluster group. The size of the rectangle is adjusted based on the log2 fold change values and the frequency of the GO term in the dataset.

CHAPTER V

IDENTIFYING THE DIRECT TAGETS OF CPVIB-1 USING CHROMATIN IMMUNOPERCIPITATION SEQUENCING

Abstract

Genetic regulation of vegetative incompatibility (*vic*) has been hypothesized to be a means of limiting the spread of hypoviruses. Compatible strains will form a stable heterokaryon, while incompatible strains will seal fused compartments that subsequently undergo programmed cell death. The transcriptional factor CPVIB-1 was found essential for the processes leading to vegetative incompatibility in *C. parasitica,* a model system for hypovirus-host interactions. In order to explore the direct targets of CPVIB-1, a detectable FLAG-tagged CPVIB-1 construct was transformed into the Δ*cpvib-1* strain. A ChIP-Seq strategy was applied to reveal the precise recognition sequence and target genes potentially regulated by the CPVIB-1 protein. We found the GAGA-repeat motif was recognized and bound specifically by the CPVIB-1 protein. We found 264 genes that CPVIB-1 targeted with function in remarkably diverse biological processes, including cell fusion, transcription, translation, autophagy, telomere maintenance, response to oxidative stress, lipid and protein biosynthesis, and carbon metabolism. From a former study, the TOR signaling pathway was demonstrated to be inactivated by both rapamycin and nutrient starvation, thus mimicking the *vic* response. Combined, our data revealed by the viable cell rate assay, hyphal growth rate assay, western blot assay, and ChIP-Seq,

indicate the connection of rapamycin, nutrient starvation, CPVIB-1 and the TOR signaling pathway in *C. parasitica*. This represents the first report that CPVIB-1 is induced by rapamycin treatment and carbon starvation treatment, that CPVIB-1 is essential for rapamycin to trigger autophagy, that CPVIB-1 is essential for both rapamycin and nutrient starvation to inhibit the hyphal growth, and that CPVIB-1 targets the TOR2 subunit directly. In conclusion, CPVIB-1 is a GAGA factor (GAF) interacting with a large collection of factors and genes that function in many different aspects of gene activity, chromosome structure and cell development and can be stimulated by rapamycin and nutrient starvation to inhibit the TOR signaling pathway.

Introduction

In the chestnut blight pathogen fungus *C. parasitica*, the genetic regulation of vegetative incompatibility (*vic*), a form of nonself allorecognition, restricts transmission of the virulence-attenuating hypoviruses [20]. Paired strains of *C. parasitica* that have allelic differences at *vic* loci display vegetative incompatibility responses, as recognized by programmed cell death leading to barrage formation [24; 95].

In *N. crassa*, VIB-1 has been highlighted as a transcriptional activator that is required for the expression of downstream effectors responsible for programmed cell death and the hyphal compartmentation triggered by the heterokaryon incompatibility (*het-c*) locus [32; 63; 66; 78]. In the plant pathogen *C. parasitica*, we have identified CPVIB-1, a putative ortholog of VIB-1, containing the same NDT80/PhoG-like DNA binding domain. In exploring the role of a putative CPVIB-1 from the model system for mycovirus-host interactions and causative agent of chestnut blight, the deletion of *cpvib-1* from the wild-type strain EP155 was performed resulting in enhanced pigmentation,

conidiation and defective vegetative incompatibility triggered by *vic4*, which indicated the role of CPVIB-1 is similar with the VIB-1 of *N. crassa* [32]. CPVIB-1 was also found to play a role in pathogenesis as well, suggesting a potential broad range of targets that this protein may influence. In this system, we predict that CPVIB-1 is part of a signaling response, which is triggered by at least some of the allelic differences that define the vegetative incompatibility system. Using a polymorphism-based comparative genomics approach, six *vic* loci in the genome were recently identified [22-24], including *vic-4*, although it is not currently known whether CPVIB-1 is responsible for signaling generated by other allelic differences.

The presence of a conserved DNA binding domain indicated CPVIB-1 may also act as a transcription factor and act to control the rate of transcription of genes by binding to specific DNA sequences in their promoter regions [96-97]. In order to further identify the downstream target genes of this potential CPVIB-1 activity, we began with comparison of large-scale transcriptome profiling of the Δ*cpvib-1* mutant strain and its isogenic wild type EP155 strain using RNA-Seq [80]. With the absence of CPVIB-1, there were 1,064 transcripts altered significantly, with 245 (23%) up-regulated and 819 (77%) down-regulated. This indicates CPVIB-1 is a key transcription regulator responsible for various biological processes, including response to oxidative stress, repressing the glucose signaling pathway, and control of the DNA, protein and lipid synthesis, all of which could contribute to affect the observed reduced virulence of the knock-out strain.

A more direct strategy to reveal the direct targets of CPVIB-1, chromatin immunoprecipitation sequencing (ChIP-Seq) [98] was applied to determine the precise

binding locations of the CPVIB-1. In 2007, Robertson G et al. first developed the ChIP-Seq method and used it to identify the DNA sequences bound by transcription factors STAT1 in interferon γ (IFNγ)-stimulated and unstimulated in human cells [99]. Since then, the ChIP-Seq method has become the most popular strategy due to its ability to identify the target DNA sequences of the protein of interest rapidly, with high efficiency and relatively low cost [100]. Here in this study, ChIP-Seq method was used to discover the DNA binding location of the transcription factor CPVIB-1 and identify the genes it targets in both rapamycin treated and untreated samples.

Rapamycin is known to be a compound with remarkable anti-fungal and immunosuppressive properties by acting with peptidyl-prolyl isomerase FPR1 to influence the Target Of Rapamycin (TOR) complex in yeast [101-102]. The TOR signaling pathway was subsequently discovered to be a conserved central controller of cell growth from yeast to humans in response to nutrient availability, stress, and growth factors [103-104]. Later in 2003, Dementhon et al. successfully linked the programmed cell death process trigged by vegetative incompatibility to the autophagy induced TOR signaling pathway in one of the model fungal species, *P. anserina* [105]. In the light of rapamycin treatment mimicking the vegetative incompatibility response, we used rapamycin in our ChIP-Seq experiment to identify the targets of CPVIB-1 that pertained specifically to the vegetative incompatibility pathway. Unexpectedly, this has revealed a relationship between CPVIB-1 and the TOR signaling pathway as well.

ChIP-Seq requires the ability to detect and recover the DNA portion of interest. Therefore, we used a FLAG-tagged CPVIB1 protein expressed in the *Δcpvib-1* mutant strain was constructed to express the detectable FLAG tagged CPVIB-1 protein. FLAG™

tag is a short peptide sequence (DYKDDDDK) that is easily and specifically recognized by the commercially available anti-FLAG antibody. This system has been widely used in western blot, immunohistochemistry, and immunoprecipitation with high specificity and protein expression, modification, purification without affecting the biological function of the tagged protein due to the small size [106-108]. With this tool validated we could then proceed with crosslinking of FLAG-CPVIB-1 protein, recovering the fixed protein-DNA complexes using the anti-FLAG antibody and then recovering the bound DNA for library preparation and analysis by high-throughput sequencing [100; 109].

With the 16 to 26 million reads generated from the FLAG-tagged CPVIB-1 samples and its rapamycin treatment samples this strategy found 275 and 357 peaks in the rapamycin untreated and treated samples, respectively. 264 (untreated) and 292 (treated) of these were found to be in the region of annotated genes using our newly prepared genome annotation (see Chapter III). We have identified GAGA-repeats that were found to be the specific binding sites of CPVIB-1 in both treated and untreated samples. Association with this GAGA motif places CPVIB-1 functionally in the GAGA factors (GAFs) family. This family of proteins has an extraordinarily diverse set of functions including the activation and repression of gene expression, nucleosome organization and remodeling, higher order chromosome architecture and mitosis [96; 110-112]. In accordance with these broad roles, CPVIB-1 was found to target genes with functions in cell fusion, transcription, translation, autophagy, telomere maintenance, response to oxidative stress, and lipid and protein biosynthesis. Furthermore, we found that the targets of CPVIB-1 protein following rapamycin treatment include the TOR complex-2

subunit and several transcriptional factors playing roles in phosphorylation, protein repair, and cell redox homeostasis.

From work in yeast, we know that both rapamycin and nutrient limitation inhibit the TOR signaling pathway, which controls the various cellular processes [103; 113]. With our results, we can now associate CPVIB-1 to the TOR signaling pathway and provide evidence for the connection between the TOR inhibitors, nutrient limitation and response to rapamycin with the CPVIB-1 protein.

Materials and Methods

Fungal spheroplast preparation

The following fungal spheroplast preparation protocol was adopted from the paper of Churchill [114]. The Δ*cpvib-1* mutant strain was previously prepared by Dr. Rong Mu and cultured in 50 ml potato dextrose broth (PDB; Difco, Sparks, MD) at room temperature on the bench for four days. After the homogenization of the fungal culture and addition of another 50 ml fresh PDB broth, the hyphal cells were expected to achieve log phase growth rate. The suspended mycelium particles were collected by filtering through sterile Miracloth® (Calbiochem) and then washed with 0.6 M $MgSO_4$ thoroughly. Harvested mycelium were then tapped dry with paper towels and transferred to a flask containing 50 ml Digestion Buffer (Table 5.1) to break down the fungal cell wall by incubating overnight on a platform shaker at 100 rpm at room temperature.

The next day, 10 ml of the spheroplast suspension was aliquoted into 50 ml polypropylene tubes and 12.5 ml cold sterile Tapping Buffer (Table 5.1) was gently overlaid onto the spheroplast suspension to form a clear interface between two layers. The tube was then centrifuged in Beckman JS13 rotor (Bucket) at 4,700 rpm at 4 ℃ for

35 minutes. The spheroplast tube was placed on ice gently and the cloudy interface containing spheroplast was collected to a new 50 ml polypropylene tube. In the same tube, 2 volumes of 1 M sorbitol solution was slowly added to the harvested spheroplast suspension with gentle mixing. The diluted spheroplast were pellet down by centrifugation in Beckman JS13 rotor at 7,000 rpm at 4 ℃ for 6 minutes.

Supernatant was removed and the spheroplast pellet was suspended in 5 ml of STC (Table 5.1) and centrifuged in Beckman JS13 rotor at 7,000 rpm at 4 ℃ for 10 minutes. Supernatant was removed and the pellet was suspended with 100 – 200 µl of the solution with 4 parts STC (Table 5.1), 1 part PTC (Table 5.1), and 0.05 parts DMSO. The re-suspended spheroplast of the Δ*cpvib-1* mutant strain were stored at -80 ℃.

Epitope FLAG-tagging of *cpvib-1* gene

The *cpvib-1* gene is 2546 nucleotides in length and encoding a protein product with 688 amino acids in length and 72 kilodaltons (KDa) in weight. The FLAG epitope tag (DYKDDDDK) was tagged to the N- terminus of the *cpvib-1* gene using the protocol described by McClean [115]. Briefly, the addition of FLAG-tag was achieved by PCR using the primer set CPVIB1-nFLAG (sense sequence: 5'-A**GCGGCCGC**ATGGATTACAAGGATGACGACGATAAGATGGCAGAGTTGAAG GAGCCGGCG -3' and anti-sense sequence: 5'-A**AAGCTT**TTACGATGTGTTCCACGAGTAGTGGCC -3'). Purified PCR products from a 1% agarose gel were cloned into the StrataClone PSC-A-amp/kan PCR cloning vector (Agilent, USA) with instruction manual for sequencing. Confirmed FLAG-tagged *cpvib-1* sequences were removed and purified from pSC-A cloning vector using the restriction sites (NotI and HindIII) designed and highlighted in the primers. The purified

FLAG-tagged *cpvib-1* DNA products were subsequently cloned into a *C. parasitica* expression vector, pCPXNBn1, containing a benomyl selection cassette.

Transformation of the FLAG-tagged *cpvib-1* gene into the Δ*cpvib-1* strain

The above construction of the expression vector and the FLAG-tagged *cpvib-1* gene was transformed into fungal spheroplast prepared in the first step of Methods and Mateirals in this Chapter [114]. At first, 200 μl of the Δ*cpvib-1* strain fungal spheroplast were quickly thawed at 37 ℃ then placed on ice. Subsequently, 5-10 μg of FLAG-tagged CPVIB-1 expression construction DNA was added to a pre-cooled 15 ml Falcon™ tube and gently mixed with 100 μl thawed spheroplast. A control was set up using autoclaved distilled water as DNA input with the same amount of spheroplast. The mixture was incubated on ice for 30 minutes and gently added 1ml PTC (Table 5.1) with a following incubation at room temperature for 25 minutes. In the end, 1ml of STC (Table 5.1) was then added and mixed gently and aliquoted to petri dishes in small droplets.

Subsequently, 12.5 ml of 45-48 ℃ Regeneration Medium (Table 5.1) was pipetted onto the droplets and swirled to mix with the transformed spheroplast thoroughly and allowed to solidify at room temperature. After 16-18 hours of incubation at room temperature, another 12.5 ml of 45-48 ℃ Regeneration Medium (Table 5.1) containing desired selective agent Benomyl (Aldrich chemistry, USA) with the final concentration as 500ng/μl was added to overlay the first Regeneration Medium layer. The plates were then kept on the bench for 5-9 days at room temperature to germinate hyphae on the top layer. Individual transformants that survived Benomyl selection were picked out to inoculate the potato dextrose agar (PDA; Difco, Sparks, MD) plates and went through sporulation

process for 7-9 days. In the end, the diluted spores were spread on the Benomyl selection PDA (final concentration as 400 ng/μl) plates for the second selection process, therefore the positive transformants were obtained from the thriving colonies from the single spores.

Expression validation of the FLAG-tagged CPVIB-1 vector construction and phenotype testing

Phenotype recovery assay

Fungal strains (EP155 wild type strain is isogenic to Δ*cpvib-1* and the positive FLAG-tagged CPVIB-1 strains) were cultured on PDA plates for 7 days to encourage hyphae growth at room temperature in the same environment. In brief, the same size plugs with mycelium from PDA plates were inoculated on the same set of PDA plates.

Vegetative incompatibility assay

The vegetative incompatibility assay was designed by Lynn Geletka to test the ability of hyphae of different strains to fuse and exchange cytoplasm [116]. The strategy of pairing the above three strains (EP155 wild type strain, its isogenic Δ*cpvib-1* strain and the positive FLAG-tagged CPVIB-1 strain) and their compatible (EU5) and incompatible strains (EU5) of *C. parasitica* was described in Chapter I. All of the above five strains were firstly revived on PDA plates for 5 days in order to obtain the fresh condition mycelium. Small plugs (4 mm^2) with mycelium were cut from all of the above five strains and paired with 3 mm separation between the two strains on BGA Medium (Table 5.1). One petri dish contains approximately 45 ml of BGA medium. The plates were then incubated in the dark for 4 days at room temperature and exposed to the light for another 2 or 3 days on the bench. Barrage formation was recorded as a manifestation of vegetative incompatibility.

Virulence assay

The three strains (EP155 wild type strain, its isogenic Δ*cpvib-1* strain and the positive FLAG-tagged CPVIB-1 strain) were started on PDA plates for 5 days again to obtain the fresh same condition mycelium. First, bark indentations (6 mm in diameter) were drilled into the American chestnut stems (*C. dentata*) approximately 20 cm apart from each other. Second, the mycelium plugs (6 mm in diameter) of three strains drilled from the PDA plates were inoculated into the indentations. Third, parafilm was wrapped around the wound with the inoculation of the fungus to prevent desiccation. At the end, the stems inoculated with three replicates of each of the three strains were kept in a clean container at room temperature for 21 days. After that, the ovate infection areas with the mycelium growth were traced on a piece of semi-transparent paper, cut out and weighed to measure the extent of the infection area. The statistics analysis of variance (ANOVA) was used to evaluate the measurements using avo package in R. Another R package ggboxplot were used to assist the analysis and display the results.

Western blot assay

The mycelial plugs from the fungal strains to be tested were inoculated on PDA plates for 5 days to obtain fresh mycelium. Plugs from these were then cultured in PDB broth on the bench, stationary and at room temperature, for four days. After homogenization and re-growth to achieve log phase fungal cells for total protein extraction, they were harvested by filtration through Miracloth, and then immediately ground into a fine powder by using a sterilized mortar and pestle [51]. If rapamycin was included, it was added at a concentration of 10 ng/ml immediately after homogenization. The total protein was then extracted using the Protein Extraction Buffer (Table 5.1) in a

ratio of 1:2 (2 ml the Protein Extraction Buffer per 1 g mycelium powder) with 50 µl protease inhibitor cocktail for yeast (Sigma, USA) per 1 g mycelium powder. The above mixture was vigorously vortexed for 15 seconds and incubated on ice for 30 mins with the above vortex and incubation one additional time. The mixture was then centrifuged at maximum speed (14,000 xg) for 15 minutes at 4°C.

The extracted total protein samples were mixed with a 1x final concentration of LDS sample buffer (Invitrogen) with the addition of β-mercaptoethanol (4 %) per manufacturer's instructions, boiled for 4 minutes and loaded on a precast 10% polyacrylamide gels (Bio-Rad Laboratories, Inc., USA) in TGS running buffer. The proteins separated in polyacrylamide gel were then transferred to an Immobilon-P membrane (Millipore, USA) using a Trans-Blot SD semi-dry electrophoretic transfer cell (Bio-Rad Laboratories, Inc., USA) using Transfer Buffer at 18 volts for 45 minutes. The membrane was briefly washed with distilled water and allowed to dry. After wetting in 100% methanol, the membrane was placed in a 5 % nonfat dry milk block (Bio-Rad Laboratories, Inc., USA) prepared in the TBST for 1 hour at room temperature or overnight at 4 °C. The membrane was briefly washed in the TBST and incubated in a 1:1000 dilution of α-FLAG monoclonal (Sigma, USA), 1:1000 dilution of α-FLAG monoclonal (Sigma, USA), 1:5000 dilution of anti-beta actin (Abcam, USA) primary antibody or 1:1000 dilution of anti-ubiquitin antibody (Cell Signaling Technology, USA) for 2 hours at room temperature. The membrane was washed three times in the TBST for 5 minutes prior to incubation in a 1:2000 dilution of HRP-Goat Anti-mouse (Bio-Rad Laboratories, Inc., USA) secondary antibody for 1 hour at room temperature. FLAG-tagged proteins were detected using Clarity Western ECL Substrate (Bio-Rad

Laboratories, Inc., USA) and imaged on a Chemi-doc Imager (Bio-Rad Laboratories, Inc., USA).

For the purpose of the detecting only ubiquitinated proteins, the membrane was then stripped with a 0.5 M NaOH solution for 15 minutes at room temperature with a gentle shaking process and rinsed three times using TBST buffer. Subsequently, the membrane was re-blocked with 5% nonfat milk and the exact same procedure above followed with the substitution of the Anti-Ub antibody (Cell Signaling Technology, Inc., USA).

Growth rate assay of nutrient starvation and rapamycin treatment on the EP155 and Δ*cpvib-1* strain

The mycelium plugs from the EP155 wild type strain and its isogenic Δ*cpvib-1* strain were inoculated on PDA plates for 5 days to obtain the same condition mycelium. The same size fresh mycelium plugs (4 mm in diameter) from above were then cultured in minimal defined media EMM (Edinburgh Minimal Media, Sunrise Science, USA), Full Medium (Add 20 g glucose and 5 g NH_4Cl per liter) plates, EMM-Low glucose (Add 5 g glucose and 5 g NH_4Cl per liter) plates, or EMM-No Nitrogen (Add 20g glucose and 0g NH_4Cl per liter) plates, all on the bench at room temperature for 7 days and all with or without 10 ng/mlrapamycin (Cayman chemical, USA). The diameters of the mycelium growth area were measured using a micrometer each day to calculate the growth rate in millimeter per day unit. The ggplot2 package of R was used to visualize the growth rate of each samples with histogram figure.

The viable cell measurement (MTT) assay of rapamycin treated mycelium

The effects of the rapamycin on cell viability of *C. parasitica* was detected by using the MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assay following the protocol from Patel [117]. Both the EP155 wild type strain and its isogenic *Δcpvib-1* strain were cultured in PDB broth for 4 days and harvested from the lower layer containing loose filamentous hyphae. Next, 50 ml of PB Buffer was used to rinse and remove residual media. Small pieces mycelium was weighed and inoculated into PB Buffer with and without 10 ng/mlRapamycin for 16 hours at room temperature. All mycelium cells were collected, rinsed with 10ml PB Buffer for 4 times and merged in the MTT Solution for the viable cells reacting with MTT to produce a purple colored formazan at 30 °C with shaking for 90 minutes. After centrifuged for 10 minutes for cell collection, the MTT Solution was removed completely and 800 μl MTT solvent was added to dissolve the purple formazan. A Gen5 Microplate Reader was used to measure the absorbance (A570) of each sample (9 samples for each stain in each treatment) [117]. All results were statistically analyzed using one-way ANOVA test using the avo package of R.

ChIP-sequencing

The mycelial plugs from the Δ*cpvib-1* strain and its isogenic positive FLAG-tagged CPVIB-1 strain were inoculated on the PDA plates for 5 days to obtain the same condition mycelium. Three biological replicates of the Δ*cpvib-1* strain and six biological replicates of the positive FLAG-tagged CPVIB-1 strain mycelium plugs were cultured in

PDB broth for 4 days and homogenized. 10 ng/ml Rapamycin for treatment was added to three of the six FLAG-tagged CPVIB-1 strain cultures.

Next day, all of the above mycelium were harvested and fixed using 1 % Formaldehyde Solution (Pierce™ 16 % Formaldehyde, Methanol-free suspended in 1X PBS, Thermo Fisher Scientific, USA) for 10 minutes on a platform shaker at 100 rpm at room temperature. Subsequently, 2.5 M glycine was used to quench the formaldehyde to a final concentration as 0.1 M for 5 minutes on the same platform shaker. After that, all mycelium was harvested on Miracloth, and then immediately washed by 1X PBS for three times.

The total protein extraction was performed using the similar protocol mentioned above but replacing the vigorous vortex step with gentle rotation for 1 hour at 4°C to minimize dissociation of the protein-DNA complex. For each 1 ml total protein solution, 3 µl of 1M $CaCl_2$ and 5 µl of micrococcal nuclease (MNase, NEB, USA) was added, and incubated for 20 minutes in a 37°C water bath with mixing every 2 minutes by inversion to digest the chromatin into small fragments. To stop the reaction and prevent over digestion, 6 µl of 0.5M EGTA (pH 8.0) was added. Next, the samples were filtered by 0.22 µm filter to remove any remaining cell debris and particulates that may interfere with protein immunoprecipitation and 140 µl of 1M NaCl was added to minimize nonspecific protein binding.

For each 1 ml nuclease digested protein-chromatin solution, 40 µl of the ANTI-FLAG M2 Magnetic Beads (Sigma, USA) was used to capture and precipitate the FLAG-tagged CPVIB-1 proteins and its bound chromatin fragments according to the

manufacturer's instruction manual. These were then eluted using a 0.1 M Glycine HCl, pH 3.0 elution protocol.

To the protein-chromatin solution, 5 µl of RNase A, DNase and protease-free 10 mg/ml (Thermo Fisher Scientific, USA) was added to a final working concentration 100ug/ml and incubated at 37°C water bath for 30 minutes. Subsequently, the solution was boiled for 10 mins and then incubated at 48°C water bath for overnight with 8µl Proteinase K (NEB, USA) to reverse the crosslinking by digesting the protein and release the DNA fragments.

ChIP DNA Clean& Concentrator Kits from Zymo Research was used to purify the DNA fragments from the above solution. 5 volumes of ChIP DNA binding buffer from the kits were used based on the DNA solution volume from the last step. 20 µl of 1X TE Buffer was used to elute the DNA from all columns of each biological replicate. The concentrations of the purified DNA fragments were examined using Qbit Fluorometer (Invitrogen, USA) and dsDNA HS Assay Kit (Invitrogen, USA).

NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB, USA) was used to prepare the ChIP DNA library using at least 1.5 ng input amounts (Table 5.2). Three steps were modified. First, in the Adapter Ligation step, instead of using 25-Fold (1:25) dilution in Tris/NaCl, pH 8.0, 50-Fold (1:50) dilution was used for all samples after the optimization in my experiment to prevent the adapter self-ligation from composing the major portion of the library. Second, in the PCR Enrichment of Adapter-ligated DNA step, both the Index Primer/i7 Primer and Universal PCR Primer/i5 Primer were reduced to half of the dose (2.5 µl instead of 5 µl) and the

NEBNext Ultra II Q5 Master Mix cut to 20 µl to avoid primer-dimer formation. Third, the number of PCR cycles for each sample was pre-optimization and set at 18.

The Illumina HiSeq2500 platform was used to sequence six samples of the ChIP-Seq DNA library from three replicates of the FLAG-tagged-CPVIB-1 and from three replicates of FLAG-tagged-CPVIB-1-Rapamycin using Illumina V4 chemistry and paired end 150 cycle reads. The sequencing was performed by Genewiz, Inc., (South Plainfield, NJ).

Bioinformatics analysis

The first step of the ChIP bioinformatics analysis was to use FastQC to perform the quality assessment of each sequencing file. Second, Trimmomatic and the adapter/primers database from the NEBNext Ultra™ II DNA Library Prep Kit was performed because the results from FastQC indicated that the existence of the low qualities of the nucleotides and the possible presence of NEB adapters and primers [83] with the SLIDINGWINDOW and MINLEN parameters setting as 4:25 and 36, separately. Third, Bowtie2 was used to align all reads using default parameters to the *C. parasitica* reference genome to generate a sequence alignment/map (SAM) format file composed of alignment information including the short reads ID, scaffold name, coordinate information, mismatch information, and the raw nucleotides of the read, among others [84]. Fourth, grep function command (grep –v "XS:i:" alignment_file.sam > alignment_file_unique.sam) was used to extract the reads that were only aligned to the genome uniquely. Fifth, Model-based Analysis of ChIP-Seq version 2 (MACS2) was used to remove the redundant reads from PCR step of ChIP DNA library preparation,

adjust read position based on fragment size distribution, calculate peak enrichment using local background normalization by running the 'callpeak' function by setting the genome size at 50 Million based on the *C. parasitica* genome assembly size, the minimum false discovery rate (FDR) q-value cutoff for peak detection at 0.01 and the fold-enrichment cutoff value at 3 to greatly eliminate the possibility of calling the false binding peaks of the CPVIB-1 protein [118]. The command used for MACS2 callpeak is listed here (macs2 callpeak -t all_biological_replicates.bam -c all_vib1_replicates.bed -g 50000000 -n outputfile_cpvib-1 -q 0.01 --fe-cutoff 3.0 -B --nomodel --trackline --verbose 3). Finally, Hypergeometric Optimization of Motif Enrichment (HOMER) version 4.10, a suite of tools for motif discovery (findMotifsGenome.pl) and annotation analysis (annotatePeaks.pl) was used to identify the recognition sequence of the CPVIB-1 protein and annotate the potential target genes of the CPVIB-1 protein with default parameters [119]. Both HOMER tools have the advantage to take the custom genome and annotation files, which can be applied to all non-model organisms rather than the built-in model organisms.

Results

Validation of functional substitution and expression of the FLAG-tagged CPVIB-1 for CPVIB-1

As discovered by Rong Mu and described in Chapter I, the Δ*cpvib-1* strain shows profuse sporulation and reduced aerial hyphal growth compared to its isogenic EP155 strain (Figure 5.1). By contrast, with the FLAG-tagged CPVIB-1 expression vector transformed into the Δ*cpvib-1* strain, the phenotype was recovered partially with increase in aerial hyphal growth and decrease in sporulation (Figure 5.1). We attribute this

difference to the use of a non-native promoter to drive expression of the FLAG tagged CPVIB-1.

The vegetative incompatibility assay was designed by culturing fungal mycelium plugs in close proximity (2~4 mm apart) on BGA Medium. Compatible colonies will fuse to create a uniform structure that becomes indistinguishable from a single colony grown alone. In *C. parasitica*, vegetative incompatibility between incompatible colonies can be seen by the production of a barrage where the colonies meet on BGA Medium. In the same study from Rong Mu in the Chapter I, the Δ*cpvib-1* strain was found to regulate the vegetative incompatibility of EP155 and EU1, which have different alleles at the *vic4* locus, when using this assay (Figure 5.2, A). However, with the deletion of *cpvib-1* from both strains, the absence of barrage in the vegetative incompatibility assay indicated they were compatible with each other (Figure 5.2, B). Based on this background, the positive FLAG-tagged CPVIB-1 strain was tested using the vegetative incompatibility assay with the EU1-Δ*cpvib-1* strain (Figure 5.2, C). The barrage formation was observed in the BGA Medium that indicated the proper operation of the FLAG-tagged CPVIB-1 protein (Figure 5.2, C).

Rong Mu also discovered that the deletion of *cpvib-1* caused a large reduction in virulence of *C. parasitica*. In this study, the virulence assay of three strains (EP155 wild type strain, Δ*cpvib-1* strain and FLAG-tagged CPVIB-1 strain) was performed on the American chestnut stems with five replicates for each strain. The cankers display of the virulence assay from these three strains indicated the virulence recovery of the FLAG-tagged CPVIB-1 strain (Figure 5.3). By utilizing the one-way ANOVA test to analyze the statistical differences of canker sizes among the three strains, the results suggested the

significantly differences of canker size between any two strains with a p-value < 0.01. The R package ggboxplot was used to display the cankers size numerical data categorized by strains (Figure 5.4). The virulence assay showed that the canker sizes of the positive FLAG-tagged CPVIB-1 strain were significantly larger than the Δ*cpvib-1* strain and close to the EP155 wild type strain. Both the phenotype assay and virulence assay results confirmed the functional substitution of FLAG-tagged CPVIB-1 for CPVIB-1 in *C. parasitica*.

In the end, western blots were used to confirm the expression and detectability of the FLAG-tagged CPVIB-1 protein by the anti-FLAG antibody with the expected 72KDa size band (Figure 5.7, Figure 5.8) An additional band higher than 75KDa protein standard (Figure 5.6, Figure 5.7, Figure 5.8) was also detected and is further described next section.

Rapamycin acts through a CPVIB-1-related pathway

Rapamycin induces cell death through CPVIB-1

In EP155 strain, with CPVIB-1 protein present, the viable cell rate in 10 ng/ml rapamycin group was 27.15 % of the control group from the MTT viable cell measurement assay. In contrast, in the Δ*cpvib-1* strain without the presence of CPVIB-1 protein, the viable cell rate was 63.83 % on average under the same rapamycin treatment (Figure 5.5). The results were statistically significant differences based on a t-test analysis (p-value < 0.001). Thus, the induced cell death rate of 10 ng/ml rapamycin was significantly decreased without the presence of protein CPVIB-1.

Rapamycin increases accumulation of the FLAG-tagged CPVIB-1

The FLAG-tagged CPVIB-1 protein accumulation level was detected by western blot for cultures grown in both EMM full Medium and 10 ng/ml rapamycin treatment/EMM full Medium (Lane 1 and 2, Figure 5.6). Rapamycin treatment increased the FLAG-tagged CPVIB-1 protein accumulation (Lane 1 and 2, Figure 5.6), consistent with the western blot assay of the positive FLAG-tagged CPVIB-1 strain cultured in PDB and 10 ng/ml rapamycin treatment PDB (Lane 1 and 2, Figure 5.7). However, the detectable protein bands were larger than the predicted size (72 KDa) and less abundant than expected when using 25 μg total protein (Figure 5.6).

CPVIB-1 is ubiquitin decorated

Based on the unexpectedly large detected protein bands from the western blot assay and the multiple positive ubiquitination sites prediction of the CPVIB-1 protein sequences using UbiSite (http://csb.cse.yzu.edu.tw/UbiSite/), it was hypothesized that CPVIB-1 could be ubiquitinated which might lead to its degradation in ubiquitin proteasome system. It is the principle mechanism for protein catabolism in cytosol and nucleus including the degradation of the transcriptional regulators like CPVIB-1 [120]. The results of the western blot assay using 80 μg total protein showed two detectable protein bands by anti-FLAG antibody, one above the 75 KDa marker that was identified previously in last step, the other below the 75 KDa marker and closer to the predicted size of 72 KDa (Lane 1 and 2, Figure 5.7). Interestingly, rapamycin treatment increased the accumulation of FLAG-tagged CPVIB-1 protein, and the upper band with the rapamycin treated samples was clearly enhanced.

Using immunoprecipitated FLAG-tagged CPVIB-1 proteins in a western blot assay with the anti-FLAG and anti-ubiquitin antibody, only the upper band was detected using anti-ubiquitin antibody, but the two band were still detected using the anti-FLAG antibody (Figure 5.8).

The connection between CPVIB-1, the nutrient starvation, and rapamycin treatment

The growth inhibition caused by rapamycin in C. parasitica is related to CPVIB-1 and the nitrogen starvation response

The growth rate of EP155 strain and its isogenic Δ*cpvib-1* strain was decreased most in the low-glucose medium with or without the presence of rapamycin (Table 5.3). The EP155 strain was found more to be sensitive to the treatment of rapamycin compared to the Δ*cpvib-1* strain (Table 5.3). The growth rate differences of two strains between treatments in EMM full medium condition suggested the growth inhibition caused by rapamycin was affected by the presence of CPVIB-1 gene (Table 5.3). This rapamycin growth inhibition affect was enhanced under the nitrogen starvation environment to the EP155 strain compared to the Δ*cpvib-1* strain (Table 5.3). However, the same phenomenon was not found in both glucose and nitrogen starvation environment (Table 5.3).

The glucose starvation environment stimulates accumulation of the FLAG-tagged CPVIB-1

Western blot assay was used to the detect the FLAG-tagged CPVIB-1 protein accumulation level from the FLAG-tagged CPVIB-1 transformed strain in both EMM full Medium and EMM-Low glucose Medium. The results suggested that the accumulation of the FLAG-tagged CPVIB-1 protein increased under the glucose

starvation condition (Lane 1 and 3 Figure 5.6). However, the nitrogen starvation treatment was found not to increase the FLAG-tagged CPVIB-1 protein accumulation level (Lane 1 and 5, Figure 5.6).

ChIP-sequencing reveals the binding recognition sequence motif and downstream genes of CPVIB-1

Quality assessment of ChIP-Seq reads

The ChIP DNA libraries were sequenced using the Illumina paired end sequencing technology, generating 32 to 83 million reads for six samples (Table 5.4) with the existence of poor quality nucleotides and NEB adapters detected using FastQC (Figure 5.9 (A)). With the application of Trimmomatic to remove the poor-quality nucleotides and NEB adapters, there were about 16 to 26 million remaining paired reads for each sample (Table 5.4). After the alignments and removal of reads that mapped to multiple locations, the single-locus mapped reads from three biological replicates of one group were combined together in BAM format to be used to call the peaks that were bound by FLAG-tagged CPVIB-1 protein.

Quality assessment of DNA fragments distribution, peak calling and regulated genes identification.

The first step of a general ChIP-Seq bioinformatics analysis is to remove the PCR duplicates. This is critical because of the low amount input DNA for preparing the library in this study [121]. However, most tools designed to perform the redundant reads removal were for sonication fragmentation projects rather than nuclease fragmentation as used in this work. However, by default MACS2 was designed to serve this goal through removing redundant reads to retain no more than one read per genomic location [118].

After this step, there were 2.69 million and 0.823 million reads retained from the FLAG-tagged-CPVIB-1 (untreated) and the FLAG-tagged-CPVIB-1-Rapamycin (treated) samples, respectively (Table 5.4).

The second step was to adjust read position based on fragment size distribution because the sequenced DNA reads from a ChIP library are often only encompassing the minimal DNA fragments spanning the protein-DNA interactions [118]. Since the sequencing technology is likely to sequence the 5' end of the both strands, the reads mapped to genome appear to be at the left or right side of the protein-DNA interaction. Therefore, the reads density around a true binding site should show a bimodal enrichment pattern, with forward strand tags enriched upstream of the binding site and reverse strand tags enriched downstream [118]. MACS2 is a model-based analysis tool designed to capture the reads that represent the ends of the fragments in a ChIP-DNA library and determine the reads that are responsible for peak calling. Both the FLAG-tagged-CPVIB-1 and the FLAG-tagged-CPVIB-1-Rapamycin samples showed the expected bimodal pattern (Figure 5.10). The d value of 128 and 112 bp for both samples represents the estimated distribution of DNA fragment size (Figure 5.10), a typical value for transcription factors and critical to extend the ChIP-Seq reads to accurately represent the original ChIP protein-DNA binding sites [118].

The third step is to calculate peak enrichment using local background normalization. As recommend by a comparison report of the tools designed for ChIP-Seq analysis, MACS2 was selected because it was designed to perform peaks calling step without the requirement of an input control sample [122]. The strategy is to model the number of reads from a genomic region using Poisson distribution with dynamic

parameter that varies along the genome to differentiate the peaks from the genome background [118]. In this study, the control samples (from strain Δ*cpvib-1*, thus a strain with no FLAG-tagged protein to recover) failed to generate sufficient DNA input for sequencing, which makes the MACS2 the best analysis tool for our project. Meanwhile, the parameters of fold-change and q-value (minimum FDR, false discovery rate) was set at 3 and 0.01 to reduce the possibility of false positive peaks calling. There were 275 and 358 peaks were called in the FLAG-tagged-CPVIB-1 and the FLAG-tagged-CPVIB-1-Rapamycin samples, separately (Table 5.5). One example of the peak that was identified as a binding site of FLAG-tagged CPVIB-1protein in PDB broth and rapamycin treatment is shown in Figure 5.11 (A). A second example of a peak only detected as a binding site of FLAG-tagged CPVIB-1 protein in PDB broth is shown in Figure 5.12 (B). A peak that was only detected as a binding site of FLAG-tagged CPVIB-1protein in rapamycin treatment is shown in Figure 5.11 (C).

Recognition sequence analysis

To reveal the FLAG-tagged CPVIB-1protein recognition sequences from the 275 and 358 peaks called in the last step, the HOMER program was used to extract sequences from the genome corresponding to the peak regions, auto-normalize the sequence bias by building the background automatically and check the enrichment of known recognition sequences of genes from its reliable library, as well as find the *de novo* recognition sequences of length less than 8 bp [119].

In the FLAG-tagged CPVIB-1 samples, the recognition sequence GAGAGAGA (Reverse Complementary: TCTCTCTC) was identified as the top candidate with a p-value of 1e-20 and exists in 68 out of 275 (24.73%) peaks (Figure 5.12(A)). The distance

between recognition sequence center and all the peak summits was calculated to determine its specific position within the peaks (Figure 5.12 (B)). A consistent feature for the transcription factors described in other studies was that the majority of recognition sequences were located within the summit region of peaks[109; 119] as was the case in the FLAG-tagged-CPVIB-1 sample data. In addition, by searching the similar recognition sequences in HOMER library, five well-studied proteins were found to bind similar recognition sequences with significant low p-values (Table 5.6) and their recognition sequences are shown in Figure 5.12 (C). The top candidate protein identified to share the similar GAGAGAGA recognition sequences is BPC6, which was identified as a transcriptional regulator involved in developmental processes in *Arabidopsis thaliana* (Figure 5.12 (C)) [112]. This GAGAGAGA recognition sequences was identified with an p-value of 1e-30 and exists in 58 out of the 275 (21.09%) peaks in the FLAG-tagged-CPVIB-1 samples (Figure 5.12 (C)). Another three known genes are FRS9, a putative transcription activator involved in regulating light control of development, BPC1, a transcriptional regulator involved in developmental processes and a GAGA-repeat binding protein that were also shown to have a highly similar GAGA recognition sequence (Figure 5.12 (C)) [123-124]. The other known protein, VRN1 is a T-repeats element binding transcriptional repressor of FLC, a major target of the vernalization pathway (Figure 5.12 (C) [125].

In the FLAG-tagged-CPVIB-1-Rapamycin samples, the recognition sequence GG[A]A[C]AGA[G]A[G]G (Reverse Complementary: CT[C]T[C]CTT[G]C[T]C[T]) was identified as the top candidate with a p-value of 1e-7 and exists in 116 out of 357 (32.49%) peaks (Figure 5.13(A)). The distance between recognition sequence center and

all the peak summits were calculated as well (Figure 5.13 (B)). The majority of this recognition sequence located within the central region of peaks in the FLAG-tagged-CPVIB-1-Rapamycin sample as well but with a diffusion distribution towards the nearby regions [109; 119]. In addition, scanning the HOMER library identified five well-studied proteins that were found to bind similar recognition sequence with significant low p-values (Table 5.7) and their motifs are shown in Figure 5.13 (C). With an p-value of 1e-7, the top candidate was identified to be KLF10, which is a transcriptional repressor that acts as an effector of transforming growth factor beta signaling in *Homo sapiens* (Figure 5.13 (C)) [126]. The same GGGGGT[C]GTGT[G]C[G]C[T] recognition sequence of KLF10(Zf) exists in 27 out of the 357 (7.56%) peaks in the FLAG-tagged-CPVIB-1-Rapamycin samples (Figure 5.13 (C)). Another three known proteins, FRS9, BPC6 and GAGA-repeat were identified in the FLAG-tagged-CPVIB-1 samples as well (Figure 5.13 (C)) [112; 123]. The other matched protein that was unique in the rapamycin treatment samples sharing a similar GAGA-repeat recognition sequence, TRL is a transcriptional activator to promote the open chromatin conformation to allow the access to other transcription factors (shown in Figure 5.13 (C)) [111].

Functional annotation of FLAG-tagged CPVIB-1 targeted genes

HOMER was used to annotate the location of the peaks in terms of important genomic features, such as transcription start site (TSS), transcription termination site (TTS), exon (Coding region), 5' UTR, 3' UTR, intronic, and intergenic using the *C. parasitica* genome and the most recent 2017-version annotation [119]. For the FLAG-tagged CPVIB-1 samples, 264 out of 275 peaks were annotated to be in the genomic features of the *C. parasitica* predicted genes (Table 5.5). The GO terms clustering

program REVIGO was used to visualize the biological processes of the targeted genes of FLAG-tagged CPVIB-1 protein (Figure 5.14). The highlighted GO term with the highest peak-score was GO:0000753 Cell morphogenesis, involved in conjugation and cellular fusion and its related gene is Ep155_U_T00007121. This is similar to the FIG1 gene from *Saccharomyces cerevisiae*, responsible for cell-cell communication, cell fusion (Figure 5.14) [127]. The second group of clustered GO terms we are interested in were the GO:0006914 autophagy and GO:0042981 regulation of apoptotic process, thus both involved in the programmed cell death pathways. The other GO term cluster groups were highlighted to represent the variety of genes involved in different biological processes, such as meiosis, metabolism, transport, transcription, translation and post-translation (Figure 5.14).

In order to investigate the expression profile of the FLAG-tagged CPVIB-1 targeted genes, the transcriptome data from Chapter IV that compared the EP155 wild type and Δ*cpvib-1* mutant strains were utilized. There were 118 out of the above 264 genes significantly altered in the Δ*cpvib-1* mutant strain suggested that about half of the observed targets of FLAG-tagged CPVIB-1 protein had their transcription process directly modulated. The GO term distribution of these 118 genes were also visualized using REVIGO [92]. The highly important GO terms mentioned above, such as, cell morphogenesis involved in conjugation, regulation of apoptosis, meiosis, metabolism, transport and transcription were found to be significantly altered in transcriptional level (Figure 5.15). In addition, several GO terms were highlighted to be significantly regulated in the Δ*cpvib-1* mutant strain included proteolysis, carbohydrate metabolism and mRNA processing, which have also been discussed in the Chapter IV (Figure 5.15).

With rapamycin treatment, 292 out of 358 peaks were annotated to be in the genomic features (Table 5.5). The majority of their GO terms were similar to the non-rapamycin treated FLAG-tagged CPVIB-1 samples, but the unique ones that were presenting with the rapamycin treatment are shown in Figure 5.16. most interestingly, these include GO:0031929 TOR signaling, which was linked to gene Ep155_U_T00000852 with a function similar to STE20 (target of rapamycin complex2 subunit) in *Schizosaccharomyces pombe* (Figure 5.16). Also, the GO terms with highest peak scores were the phosphorylation and phosphorelay signal transduction systems. In addition, the protein repair, sterol metabolism, exocytosis and cell redox homeostasis were highlighted after rapamycin but not were not present in the untreated analysis (Figure 5.16). Of the 292 genes, 121 were significantly regulated at the transcriptional level by CPVIB-1 from the transcriptome comparison data. In the GO terms cluster distribution plot, they were all overlapped with the GO terms from the targets the FLAG-tagged CPVIB-1 (untreated) samples (Figure 5.17).

In the 264 genes targeted by FLAG-tagged CPVIB-1 and 292 genes targeted by FLAG-tagged CPVIB-1 with rapamycin treatment, 21 of them were same genes from them and the others were unique based on the treatment (Table 5.5). All of them are listed with their UniProt ortholog protein names, the log2 fold change value in the transcriptional level in the absence of CPVIB-1, and the GO terms and biological processes they involved in (Table 5.8). Among them, the apoptotic regulation gene (YBR241C), the glucose metabolism gene (cfp, pyruvate decarboxylase), the oxidation-reduction process (mutM), the mRNA degradation process (dom34, RNA surveillance) and metabolism (SPBC725.05C) are all plausibly related to the phenotype shift observed

in the mutant *Δcpvib-1* strain, such as the reduced hyphal growth, the reduced virulence and reduced programmed cell death. Moreover, a transcriptional regulator named SIM1 is a protein involved in the control of the apoptosis-like cell death and telomere length homeostasis with CDC13 in *S. cerevisiae* [128].

Discussion

As a preparative step to revealing the targets of CPVIB-1 in this study, the FLAG-tagged CPVIB-1 was expressed and shown to functionally complement the *C. parasitica* Δ*cpvib-1* mutant strain, and to be recoverable by immunoprecipitation. During this work, CPVIB-1 was found to be decorated by ubiquitin which might lead to the degradation of CPVIB-1 through the ubiquitin-proteasome system (UPS). UPS is a universal regulation system shown to control the localization, abundance and activity of transcription factors [129]. As one of the principal mechanism for protein catabolism, the UPS regulates the abundance of the transcription activators by covalently tagging the proteins with multiple ubiquitin molecules (Conjugation), and subsequently degrading the tagged proteins via the proteasome activity (Degradation) [120; 130]. Based on the bands migration compared to the protein molecular weight standards, we estimated that the mass of ubiquitin modified FLAG-tagged CPVIB-1 protein to be approximately 22 KDa larger than the predicted size and much more abundant led to the hypothesis that the CPVIB-1 protein was associated with multiple ubiquitin modifications at one or various sites or other types of post-translation modification.

CPVIB-1 was found to be induced by rapamycin and carbon starvation treatment. In the study of vegetative incompatibility system in fungus *P. anserina*, rapamycin was

discovered to be able to mimic the incompatibility reaction by targeting the TOR (target of rapamycin) signaling pathway, which has the TOR protein kinase (conserved from humans to yeast) to control cell growth in response to nutrient availability [104-105]. Generally, rapamycin inhibits the TOR signaling pathway leading to the autophagy, expression of starvation-induced genes and inhibition of translation in yeast and other organisms [102; 105]. Therefore, rapamycin was hypothesized to be a promoter for the expression of CPVIB-1 to induce programmed cell death. To test this hypothesis, the viable cell number in the rapamycin treated and untreated were measured using the MTT assay, the hyphal growth rate in the rapamycin treated, nutrient starvation treated and untreated were measured, and the accumulation level of FLAG-tagged CPVIB-1 in the rapamycin treated, nutrient starvation treated and untreated samples was detected.

The viable cell number in rapamycin treated mycelium was significantly reduced compared to the control indicating the same inhibition mechanism of cell growth through TOR pathway in *C. parasitica* as was observed for *P. anserina*. Second, without the presence of CPVIB-1, the ability of rapamycin to trigger cell death was reduced, indicating the relationship between the CPVIB-1 and rapamycin action. Third, the hyphal growth rate inhibition of the rapamycin treatment was reduced without the presence of CPVIB-1 compared to EP155 strain. Lastly, the accumulation level of FLAG-tagged CPVIB-1 protein in the rapamycin treated tissue was clearly increased compared to the control confirming that CPVIB-1 is induced by rapamycin treatment. Under the nutrient starvation environment, the TOR signaling pathway was inactivated leading to the similar outcomes for rapamycin treatment in *P. anserina* [105]. Therefore, in *C. parasitica*, the connections of nutrient starvation, rapamycin and CPVIB-1 was tested using the growth

rate assay and the western blot assay. The growth rate under the glucose starvation but not nitrogen starvation was significantly reduced compared to the full defined medium. Similarly, without the presence of CPVIB-1, the growth rate inhibition of rapamycin treatment under either of the nutrient starvation was also reduced indicating both rapamycin and nutrient starvation inactivate the TOR signaling pathway. Furthermore, while nitrogen starvation did not induce CPVIB-1 expression, the glucose starvation was found to increase the accumulation of CPVIB-1 similar to the rapamycin treatment. The related actions of CPVIB-1 and rapamycin present the clear association of regulatory mechanisms controlling the TOR signaling pathway.

With current technology, ChIP-Seq is the most widely used strategy to explore the protein-DNA interaction sites, and is suitable for the goal of this study to identify the targets of a transcriptional factor [131]. In this study, the ChIP DNA library prepared from the FLAG-tagged CPVIB-1 samples was sequenced for this purpose. Meanwhile, the FLAG-tagged CPVIB-1 rapamycin treatment samples were prepared to induce the expression of CPVIB-1 and sequenced as well to explore the possibility of the treatment leading to differential targets and potentially reveal the regulatory relationship of TOR and CPVIB-1. Unfortunately, it was impossible to obtain enough background input DNA for the negative control probably due to the high specificity of the ANTI-FLAG antibody under the 160 mM sodium salt lysis solution, which is recommended for the high specificity binding. As for the quality of the sequencing, 16 to 26 million reads were obtained for each biological replicate after the stringent trimming process (Table 5.4). ENCODE ChIP-Seq guideline recommended that 20 million reads are suitable for mammalian studies and 4 million for the *Caenorhabditis elegans* [131]. Therefore, there

were ample high-quality DNA reads for our project based on the relatively small genome size [131]. Because of the absence of the background input negative control, the peak-calling process was set with the extremely harsh parameters to eliminate false discovery. For both the FLAG-tagged CPVIB1 sample and its rapamycin treated sample, the GA-rich element was revealed to be the recognition sequence of CPVIB-1 protein. By searching the similar motif binding genes against the HOMER database, several transcriptional regulators with the same GA-rich motif binding feature were identified. They were BPC1, BPC6, FRS9 and VRN1, three GAGA-biding transcriptional regulators in *Arabidopsis* plus the TRL and GAGA-repeats factors, two transcriptional regulators in *Drosophila* [111-112; 123-125].

Therefore, we conclude that CPVIB-1 is functionally a GAGA-repeats binding transcriptional regulator (GAF) in *C. parasitica* that functions in remarkably diverse range of regulatory contexts, including activation/repression of genes' transcription, mitosis maintenance, cell development, and autophagy, cell-cell communication, virulence, nucleosome assembly, etc. In other systems, GAF transcriptional regulators were found to be responsible for the activation and silencing of gene expression by specific binding to the promoter regions of various genes, including homeotic genes during developmental stages [96]. In this study, CPVIB-1 was found to target 275 genes distributed in various biological processes and significantly altered about half of them at the transcription level. These processes included mitosis, biosynthesis, metabolism, autophagy, cell fusion, and translation, which are consistent with the roles found from former studies of GAFs in other systems (Figure 5.15) [112; 124; 132-134].

CPVIB-1 targets at least six important regulators. SIM1 is involved in the control of the apoptosis-like cell death and telomere length homeostasis [128], which correlates with the function of CPVIB-1 to maintain pathogenesis that was observed in both Chapter I and IV. Epithelial mesenchymal transition (EMT) plays roles in altering cells to a mesenchymal cell phenotype [135]. Proteasome regulator is an essential component of the ubiquitin-proteasome system [130]. GTPase signaling regulator activates GTPase signaling pathway and involves in the virulence [136-137]. CRZ1 regulator is essential for the pathogenicity by regulating the calcium signaling and high-osmolarity glycerol response (HOG) signaling pathway [109; 138]. In the end, SAP30 regulator was discovered to be a component of histone deacetylase complex, which involved in telomere maintenance and the infectious growth of rice blast fungus [139].

Furthermore, GAFs have been found to act as more than a typical transcriptional factor but also the genome-scale transcriptional regulator by facilitating the formation and maintenance of nucleosome free regions of the chromatin to expose the promoter sequences for the access of other transcriptional factors [111]. In this study, the target genes of CPVIB-1 include the DNA helicase for the mitosis, telomere maintenance and the decomposition of the chromatin complex and the activators for the transcription of RNA polymerase II promoter for the promotion of the transcription process [140-142] (Figure 5.15), all of which would fit with this role. In conclusion, CPVIB-1 is a GAGA-motif biding transcriptional regulator (GAFs) in *C. parasitica*. This is the first confirmed GAGA repeats binding regulator identified in fungi that functions in the regulating of various downstream transcription factors and downstream genes involved in diverse biological processes including homeotic genes as well as in the regulation of the genome-

wide transcription process by promoting the chromatin decomposition for releasing the DNA segments for transcription polymerase and other factors.

By mimicking the vegetative incompatibility process using rapamycin, the recognition sequence of CPVIB-1 was shifted slightly to target genes involved in the autophagy, phosphorylation, sporulation, cell redox hemostasis, protein glycosylation, exocytosis, and the TOR signaling pathway. With the treatment by rapamycin, the FLAG-tagged CPVIB-1 protein targets 357 genes and only 33.9% of them were significantly altered in the transcriptome profile of the CPVIB-1 deletion strain, but there were a few new noteworthy targets.

The apoptosis-inducing factor 2 gene encodes an oxidoreductase, which plays a role in mediating the apoptotic cell death by binding to a sequence-specific DNA site [143]. The LAE secondary metabolism regulator coordinates response to light in *Aspergillus* and regulates T-toxin production, virulence, oxidative stress response and cell development of maize pathogen *Cochliobolus* [144]. The cAMP-independent regulator PAC2 was found to be required for sporulation, regulating the meiotic cell cycle and negatively regulating cell fusion [145]. Furthermore, the identified *C. parasitica* version of cAMP-independent regulator PAC2 was significantly up-regulated in the CPVIB-1-lacking strain. Lastly, the aspartic protease PEP1 gene, encodes a secreted aspartic endopeptidase that contributes to the virulence [146]. In conclusion, there is a remarkable confluence of the phenotypic characteristics of a strain lacking CPVIB-1 and the specific gene targets that this procedure has revealed that are involved in programmed cell death, virulence, meiosis, cell fusion, and sporulation progress.

Intriguingly, with the treatment of rapamycin, CPVIB-1 was found to target the TOR2 subunit STE20, one component of the TOR Complex [147]. The TOR signaling pathway includes two segments, TOR1 and TOR2 [102]. TOR1 is the interaction site for rapamycin and is responsible for regulating the sensing of stress, growth factors, nutrients to regulate the protein synthesis, lipid and nucleotide synthesis, and autophagy. TOR2 was found to be insensitive to rapamycin but sensitive to growth factors that regulated survival and proliferation [102]. Meanwhile, two other genes, the hybrid signal transduction histidine kinase K and acetate kinase, were targeted by CPVIB-1 during the rapamycin treatment with the highest peak scores and they are involved in the phosphorylation and the phosphorelay signal transduction biological processes. With the discovery that the TOR contained a C-term kinase domain, it was found to play a central regulator role in an abundant and diverse set of genes which are important for cell growth by reacting with related kinase family proteins [102]. In conclusion, in the presence of rapamycin CPVIB-1 targets the TOR2 and other kinases which are involved in the TOR signaling pathway.

Ultimately, we have shown that both the carbon starvation and rapamycin inhibit the TOR signaling pathway leading to the autophagy and inhibition of cell growth, consistent with other systems. However, CPVIB-1 was found to target TOR2 expression and be induced by rapamycin and carbon starvation treatment. Therefore, we propose a model such that rapamycin or a nutrient starvation signal targeting TOR1 results in changes to CPVIB-1 expression. This, in turn, affects TOR2-regulated pathways by perturbing important members of the complex. Thus, we propose a novel regulatory

mechanism whereby CPVIB-1 acts as a mediator to link the separate regulatory regimes of TOR1 and TOR2-dependent pathways.

Tables and Figures

Table 5.1 Solution recipes used in this chapter.

Solution Name	Components	Sterilization protocol
Osmotic Medium	1.2 M $MgSO_4$; 10 mM NaH_2PO_4; adjust pH to 5.8 with 0.5 M Na_2HPO_4	filter sterilization
Digestion Buffer	1 % β-glucuronidase [Sigma G-7770]; 0.2 % Lysing enzyme [Sigma L-1412]; 0.6 % Bovine Serum Albumin [Fisher BP 1605]; 1.5 % Vinoflow™	filter sterilization
Trapping Buffer	0.4 M Sorbitol; 100 mM Tris-HCl pH 7.0	autoclave sterilization
STC	1 M Sorbitol; 100 mM $CaCl_2$; 100 mM Tris-HCl pH 8.0	autoclave sterilization
PTC	40 % Polyethylene glycol 4000 MW; 100 mM Tris-HCl pH 8.0; 100 mM $CaCl_2$	autoclave sterilization
Regeneration Medium	1 M sucrose; 0.1 % Yeast Extract; 0.1 % Casein Hydrolysate	autoclave sterilization
BGA Medium	2.4 % PDA; 0.7 % malt extract agar; 0.2 % yeast extract; 0.08 % tannic acid; 0.005 % mg Bromocresol Green; 0.06 % drops Tween 20; 2 % BD Bacto™ Agar	autoclave sterilization
Protein Extraction Buffer	50 mM Tris-HCl pH 8.0; 20 mM NaCl; 1 % Triton-X100; 0.1 % CHAPS; 0.1 % NP-40	filter sterilization
TGS Running Buffer	25 mM Tris; 192 mM Glycine;	No sterilization

	0.1 % SDS, pH 8.3	
TBST	20 mM Tris; 150 mM NaCl; 0.05 % Tween 20, pH 7.6	No sterilization
Transfer Buffer	25 mM Tris; 192 mM glycin; 20 % Methanol	No sterilization
PB Buffer	75.4 mM Na_2HPO_4; 24.6 mM NaH_2PO_4	autoclave sterilization
MTT Solution	50 mg Thiazolyl Blue Tetrazolium Bromide; dissolved in 10 ml PB Buffer	filter sterilization
MTT Solvent	4 mM HCl; 0.1 % NP-40; dissolved in isopropanol	filter sterilization
1X TE Buffer	10 mM Tris-Hcl ph 8.0; 1 mM EDTA	filter sterilization

Table 5.2 The mass of DNA extracted from ChIP, the concentration of ChIP libraries, and the DNA fragment size of ChIP libraries.

Sample name	Mass of Purified DNA (ng)	Concentration of purified DNA library (ng/ul)	DNA fragments size
Δ*cpvib-1*-rep1	[1]Too little to be detected	0.6	[2]Too low
Δ*cpvib-1*-rep2	[1]Too little to be detected	1.22	[2]Too low
Δ*cpvib-1*-rep3	[1]Too little to be detected	0.4	[2]Too low
FLAG-tagged-CPVIB-1-rep1	1.5	6.16	273
FLAG-tagged-CPVIB-1-rep2	2.0	12.8	272
FLAG-tagged-CPVIB-1-rep3	2.7	18.1	286
FLAG-tagged-CPVIB-1-Rapamycin-rep1	2.1	7.74	275
FLAG-tagged-CPVIB-1- Rapamycin-rep2	2.5	10.5	280
FLAG-tagged-CPVIB-1- Rapamycin-rep3	3.4	22.8	317

[1] Too little to be detected indicated the concentration of the DNA samples lower than the lowest threshold of the Qbit Fluorometer.
[2] Too low indicated the concentration of the DNA library was lower than or close to 1 ng/μl that the DNA chip of Agilent 2100 Bioanalyzer® was not able to detect it.

Table 5.3 The growth rate of the EP155 and its Δ*cpvib-1*strain on various media with and without rapamycin treatment.

Medium condition	Growth rate (mm/day)	
	EP155	Δ*cpvib-1*
EMM	9.27±0.600	8.70±0.263
EMM-Low glucose	7.98±1.135	7.01±0.139
EMM-No Nitrogen	8.82±0.416	8.01±0.482
EMM-Low glucose-No Nitrogen	8.57±0.269	8.40±0.225
EMM-10 ng/ml rapamycin	5.49±0.990	5.71±0.266
EMM-Low glucose-10 ng/ml rapamycin	5.52±0.029	5.26±0.212
EMM-No Nitrogen-10 ng/ml rapamycin	5.70±0.137	6.36±0.322
EMM-Low glucose-No Nitrogen-10 ng/ml rapamycin	5.57±0.284	5.42±0.289

Growth rate (mm/day) was calculated from the growth diameter of fungus each day in solid media with three biological replicates. The numbers represent the growth rate and the standard deviation within the three biological replicates.

Table 5.4 Sequenced reads numbers from ChIP-Seq library.

Samples	Original reads number (Million)	Remained reads number (Million)	Total reads uniquely-mapped and response to call peaks (Million)	Non-redundant reads response to call peaks (Million)
FLAG-tagged-CPVIB-1-rep1	43	19	15.6	2.69
FLAG-tagged-CPVIB-1-rep2	83	26		
FLAG-tagged-CPVIB-1-rep3	43	20		
FLAG-tagged-CPVIB-1-Rapamycin-rep1	42	16	18.0	0.823
FLAG-tagged-CPVIB-1- Rapamycin-rep2	49	24		
FLAG-tagged-CPVIB-1- Rapamycin-rep3	32	16		

Table 5.5 The results of called peaks from MACS2.

Samples	Peaks Number				
	Total	Annotated to link to genes	Unique in samples	Present in transcriptome significantly regulated genes	Unique in transcriptome significantly regulated genes
FLAG-tagged-CPVIB-1	275	264	243	118	106
FLAG-tagged-CPVIB-1-Rapamycin	357	292	261	121	109

Table 5.6 The top five genes that recognize similar sequences with FLAG-tagged-CPVIB-1 protein from HOMER database.

Motif Name	Consensus	P-value	# of Target Sequences with Motif (of 275)	% of Target Sequences with Motif	% of Background Sequences with Motif
PC6(BBRBPC)/col-BPC6-DAP-Seq	YTYTCTCTCTCTCTA	1.00E-30	58	21.09%	3.00%
FRS9(ND)/col-FRS9-DAP-Seq	RGAGAGAGAAAG	1.00E-18	68	24.73%	7.43%
BPC1(BBRBPC)/colamp-BPC1-DAP-Seq	GARGAGAGAGAA	1.00E-11	68	24.73%	10.07%
VRN1(ABI3VP1)/col-VRN1-DAP-Seq	TTTTTTTTTT	1.00E-10	38	13.82%	3.98%
GAGA-repeat/Arabidopsis-Promoters	CTCTCTCTCY	1.00E-08	79	28.73%	14.75%

Table 5.7 The top five genes that recognize similar sequences with FLAG-tagged-CPVIB-1 protein with the treatment of rapamycin from HOMER database.

Motif Name	Consensus	P-value	# of Target Sequences with Motif (of 357)	% of Target Sequences with Motif	% of Background Sequences with Motif
KLF10(Zf)/HEK293-KLF10	GGGGGTGTGTCC	1.00E-05	27	7.56%	2.64%
FRS9(ND)/col-FRS9-DAP-Seq	RGAGAGAGAAAG	1.00E-05	42	11.76%	5.71%
BPC6(BBRBPC)/col-BPC6-DAP-Seq	YTYTCTCTCTCTCTA	1.00E-04	22	6.16%	2.26%
Trl(Zf)/S2-GAGAfactor-ChIP-Seq	RGAGAGAG	1.00E-04	131	36.69%	27.13%
GAGA-repeat/Arabidopsis-Promoters	CTCTCTCTCY	1.00E-03	58	16.25%	10.18%

Table 5.8 The 21 target genes of FLAG-tagged CPVIB-1 protein with and without rapamycin treatment.

Nearest PromoterID	Gene name	log2 fold change	GO terms	Involved biological processes
Ep155_U_T00001309_1	Similar to Slc38a3: Sodium-coupled neutral amino acid transporter 3 (Rattus norvegicus)	0.76		
Ep155_U_T00006432_1	Similar to DDB_G0272472: Calponin homology domain-containing protein DDB_G0272472 (Dictyostelium discoideum)	-0.57		
Ep155_U_T00002163_1	Similar to SPAC10F6.14c: ABC1 family protein C10F6.14c (Schizosaccharomyces pombe (strain 972 / ATCC 24843))	-0.47		
Ep155_U_T00002322_1	Similar to rplV: 50S ribosomal protein L22 (Arcobacter butzleri (strain RM4018))	4.12		
Ep155_U_T00003573_1	Similar to cfp: Pyruvate decarboxylase (Neurospora crassa (strain ATCC 24698 / 74-OR23-1A / CBS 708.71 / DSM 1257 / FGSC 987))	0.70	GO:0000287, GO:0003824, GO:0016831, GO:0030976	Glucose metabolism
Ep155_U_T00004752_1	Similar to blr3397: Nitrilase blr3397 (Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110))	0.58	GO:0006807, GO:0016810	Nitrogen compound metabolic process
Ep155_U_T00008921_1	Similar to YBR241C: Probable metabolite transport protein YBR241C (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	0.46	GO:0005215, GO:0016020, GO:0016021, GO:0022857, GO:0022891, GO:0042981, GO:0055085	Regulation apoptotic process, Transmembrane transport

Table 5.8 (continued)

Ep155_U_T00003188_1	Similar to Zan: Zonadhesin (Mus musculus)	1.12		
Ep155_U_T00002617_1	Similar to SEC31B: Protein transport protein SEC31 homolog B (Arabidopsis thaliana)	-0.31		
Ep155_U_T00009404_1	Similar to mutM: Formamidopyrimidine-DNA glycosylase (Lactococcus lactis subsp. cremoris)	1.51	GO:0016491, GO:0055114	Oxidation-reduction process
Ep155_U_T00002128_1	Similar to GPI19: Phosphatidylinositol N-acetylglucosaminyltransferase subunit GPI19 (Ashbya gossypii (strain ATCC 10895 / CBS 109.51 / FGSC 9923 / NRRL Y-1056))	-2.82		
Ep155_U_T00003820_1	Similar to CAH: Cyanamide hydratase (Myrothecium verrucaria)	/		
Ep155_U_T00005122_1	Similar to AAH1: Adenine deaminase (Gibberella zeae (strain PH-1 / ATCC MYA-4620 / FGSC 9075 / NRRL 31084))	/	GO:0000034, GO:0006146, GO:0009168, GO:0019239	Purine ribonucleoside monophosphate biosynthetic process, Adenine catabolic process
Ep155_U_T00000385_1	Similar to STM1: Suppressor protein STM1 (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	/		
Ep155_U_T00009393_1	Protein of unknown function	/		
Ep155_U_T00001837_1	Similar to At2g17033: Pentatricopeptide repeat-containing protein At2g17033 (Arabidopsis thaliana)	/		
Ep155_U_T00001179_1	Similar to dom34: Protein dom34 (Schizosaccharomyces pombe (strain 972 / ATCC 24843))	/	GO:0070481, GO:0070966, GO:0071025	RNA surveillance, Nuclear-transcribed mRNA catabolic process, no-go decay
Ep155_U_T00002100_1	Similar to nuf2: Kinetochore protein Nuf2 (Danio rerio)	/		
Ep155_U_T00002361_1	Similar to SPBC725.05c: Uncharacterized pyrophosphatase/phosphodiesterase C725.05c (Schizosaccharomyces pombe (strain 972 / ATCC 24843))	/	GO:0003824, GO:0008152	Metabolic process
Ep155_U_T00005230_1	Similar to AF_2170: Kelch domain-containing protein AF_2170 (Archaeoglobus fulgidus (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126))	/		
Ep155_U_T00001196_1	Similar to tcmP: Tetracenomycin polyketide synthesis O-methyltransferase TcmP (Streptomyces glaucescens)	/	GO:0008168, GO:0032259	Methylation

Nearest promoterID is the predicted gene ID from 2017-version annotation that the peaks located in their gene promoter regions. Log2 fold change is the transcription level changes in the Δ*cpvib-1* strain. ('/' indicates the transcription level were not significantly regulated).

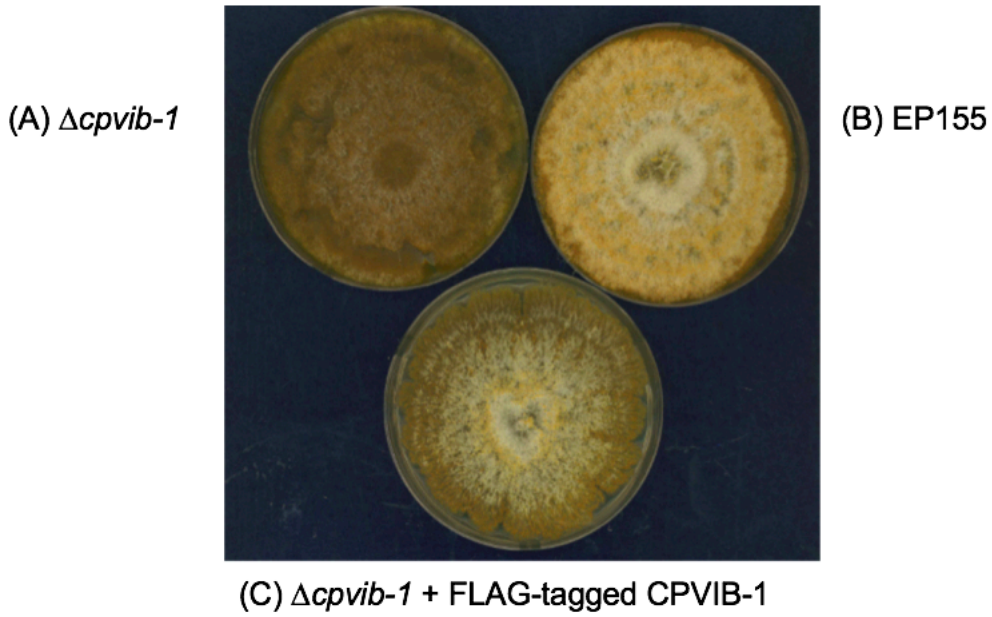


Figure 5.1 Phenotype recovery assay from the Δ*cpvib-1* strain with the FLAG-tagged CPVIB-1 expression vector.

(A) The representative phenotype of Δ*cpvib-1* strain with profuse sporulation and reduced aerial hyphal growth compared to the representative phenotype of its isogenic EP155 strain (B).
(C) The representative phenotype of the FLAG-tagged CPVIB-1 strain with the expression vector transformed into the Δ*cpvib-1* strain (C) was partially recovered compared to the representative phenotype of its isogenic EP155 strain (B).

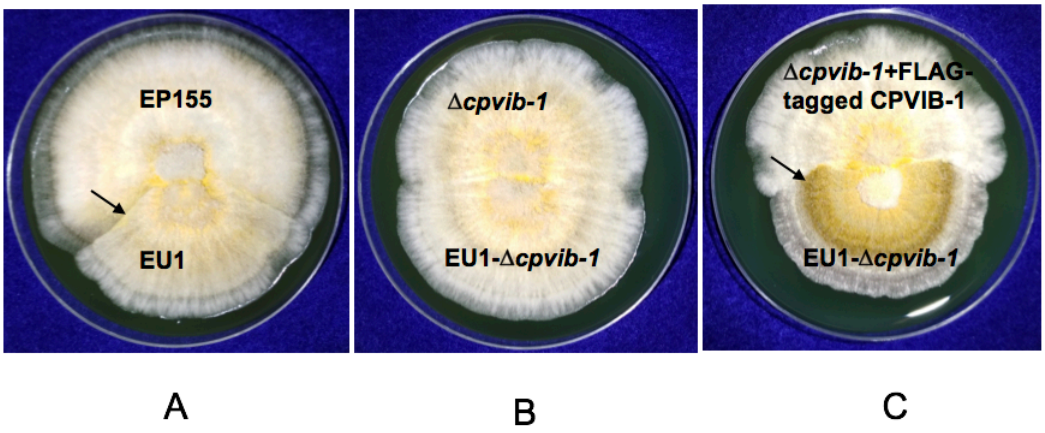


Figure 5.2 Vegetative incompatibility assay of *C. parasitica*.

(A) EP155 wild type demonstrated incompatibility with EU1. (B) Deletion of *cpvib-1* from both EP155 and EU1 converted them to the compatible fusion with each other. (C) The FALG-tagged CPVIB-1 expression vector transformed in the Δ*cpvib-1* strain recovered the incompatibility with EU1-Δ*cpvib-1* strain.

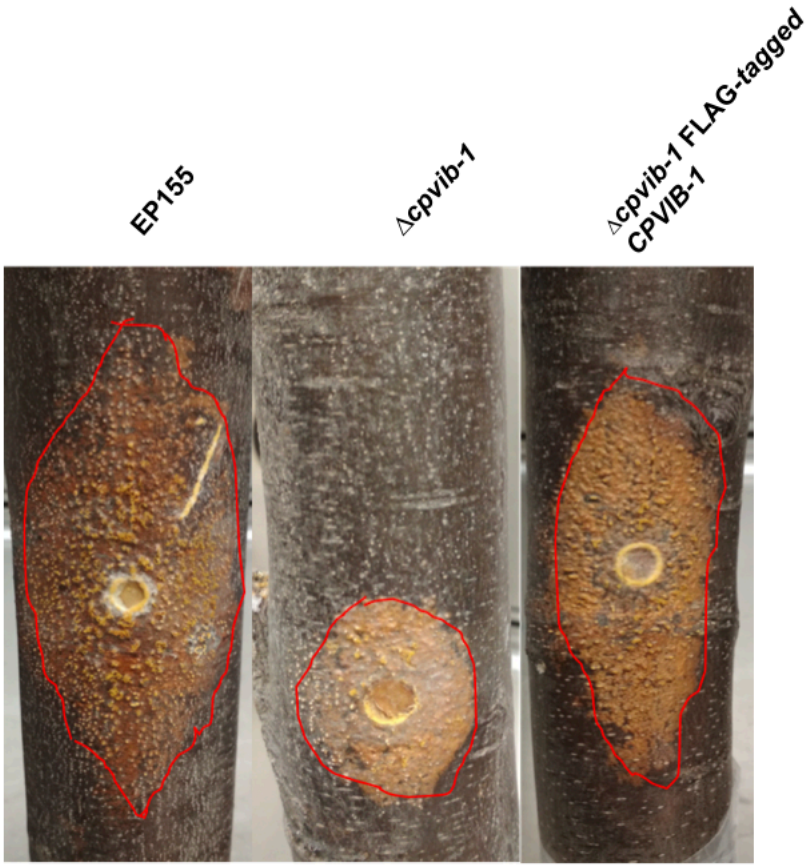


Figure 5.3 Cankers display of the virulence assay with various *C. parasitica* strains infecting the American chestnut stems.

The representative cankers on dormant chestnut stems were caused by *C. parasitica* after 21 days. Canker sizes were highlighted with red lines for each strain.

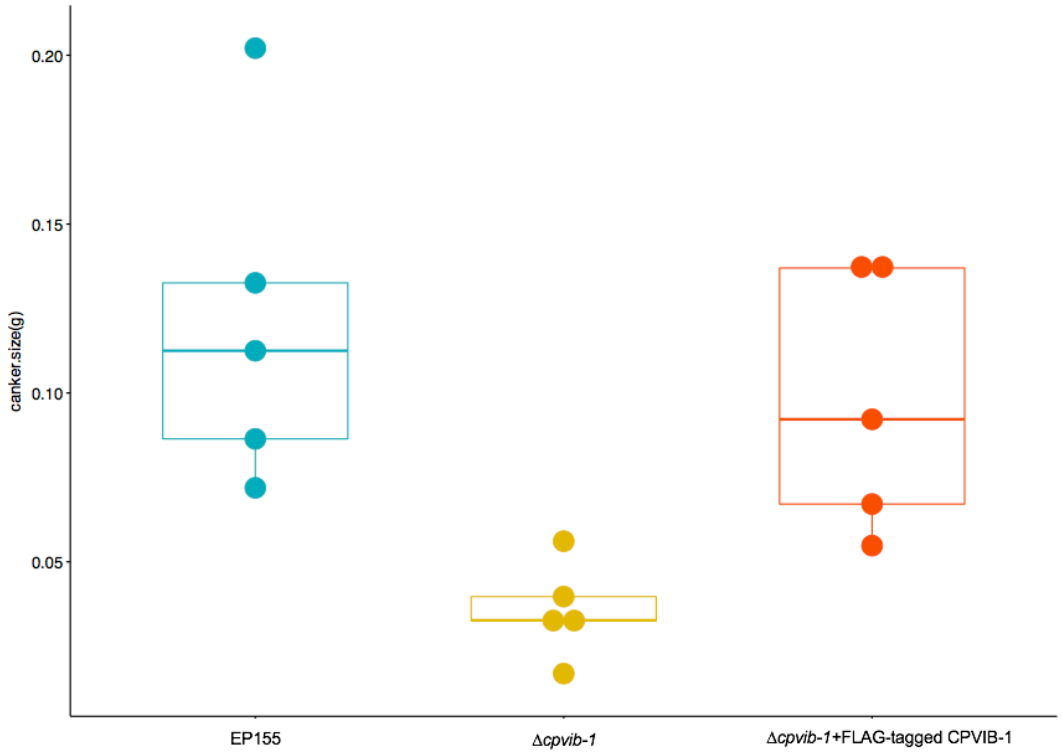


Figure 5.4 Cankers size analysis of the virulence assay with various *C. parasitica* strains infecting the American chestnut stems.

Boxplot represented the cankers size from virulence assay. Each dot represented the canker size of one sample, the horizontal line in the middle of the box represented the mean number of the cankers size in each strain and the vertical line at the bottom of the box represented of the standard deviation.

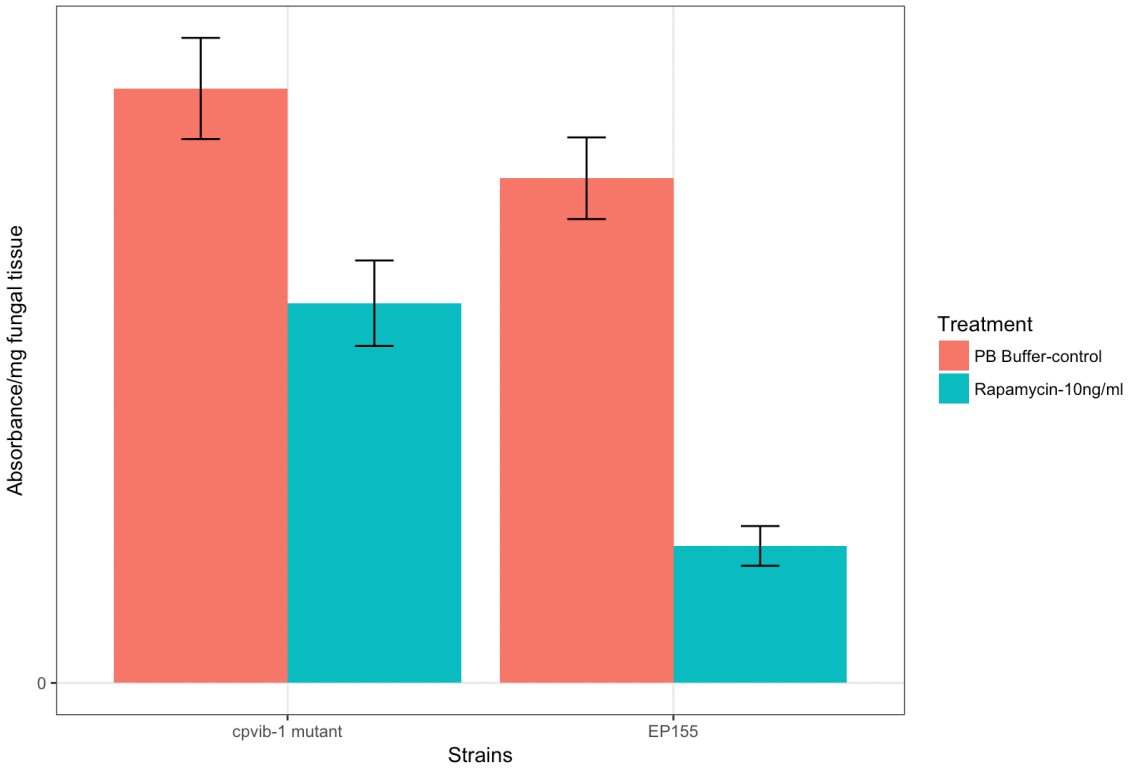


Figure 5.5 The *cpvib-1* mutant strain showed less viability reduction when treated with rapamycin. The proportion of viable cells were measured using the MTT viability assay with and without 10 ng/ml rapamycin treatment.

The y axis is the absorbance/mg fungal tissue obtained from the spectrophotometer at 570 nm wavelength. The x axis is the rapamycin treatment (Green bars) and PB buffer control (Red bars) within two strains. The black line at the top of each bar is the standard deviation among three biological replicates, each with three technical replicates.

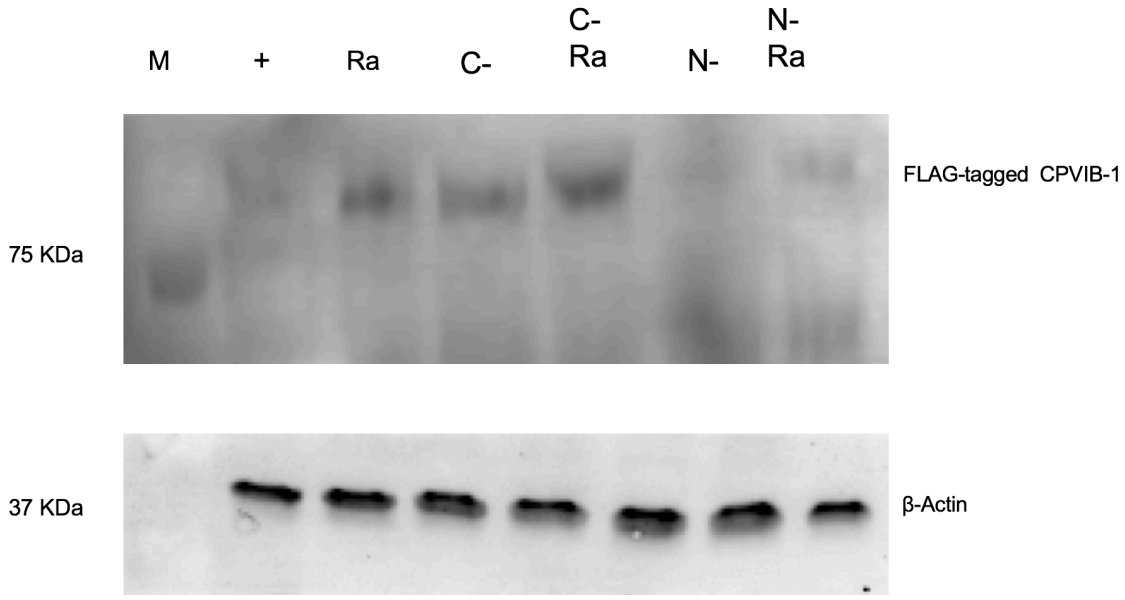


Figure 5.6 The representative image from western blot assay of FLAG-tagged CPVIB-1 proteins from the nutrient starvation and rapamycin treatment.

Lane M, 15 µl of Bio-Rad protein ladder,
lane 1, 25 µg total proteins from EMM Medium culture,
lane 2, 25 µg total proteins from EMM Medium culture with 10 ng/ml rapamycin,
lane 3, 25 µg total proteins from EMM-Low glucose culture,
lane 4, 25 µg total proteins from EMM-Low glucose with 10 ng/ml rapamycin,
lane 5, 25 µg total proteins from EMM-No Nitrogen culture,
lane 6, 25 µg total proteins from EMM-No Nitrogen with 10 ng/ml rapamycin.

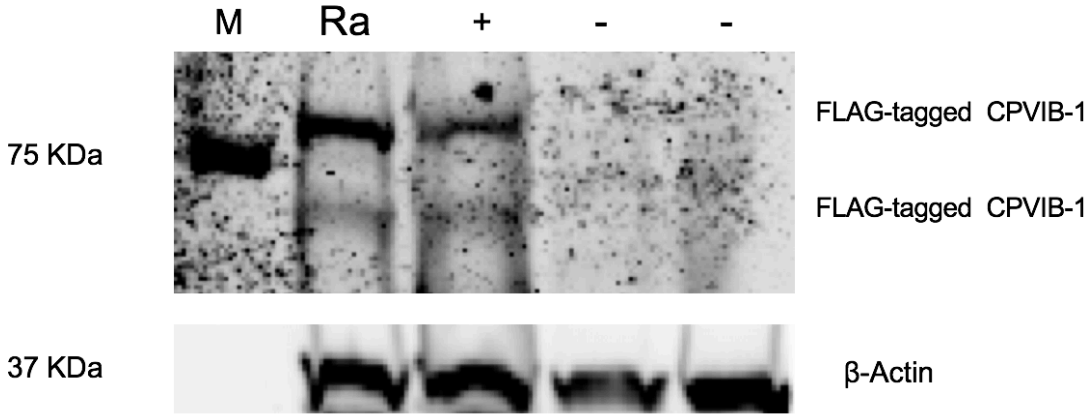


Figure 5.7 The representative image from western blot assay of the FLAG-tagged CPVIB-1 protein using anti-FLAG antibody.

Lane M, 15 μl of Bio-Rad protein ladder,
lane 1, 80 μg total proteins from the positive FLAG-tagged CPVIB-1 strain treated with 10 ng/ml rapamycin,
lane 2, 80 μg total proteins from the positive FLAG-tagged CPVIB-1 strain (positive control),
lane 3, 80 μg total proteins from EP155 strain (negative control),
lane 4, 80 μg total proteins from Δ*cpvib-1* strain (negative control).

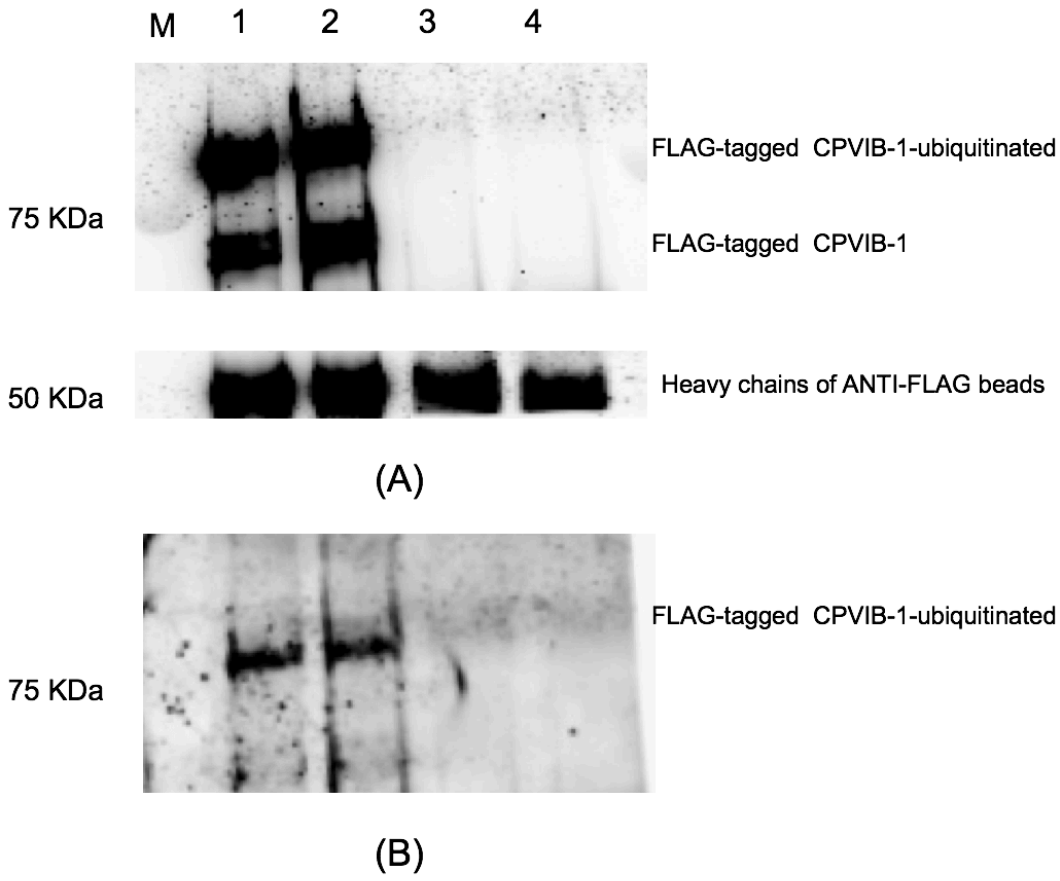
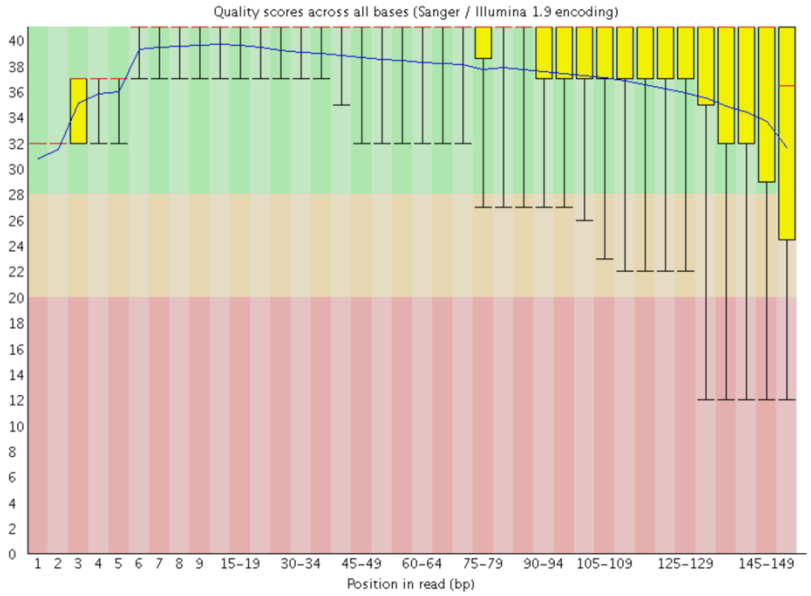
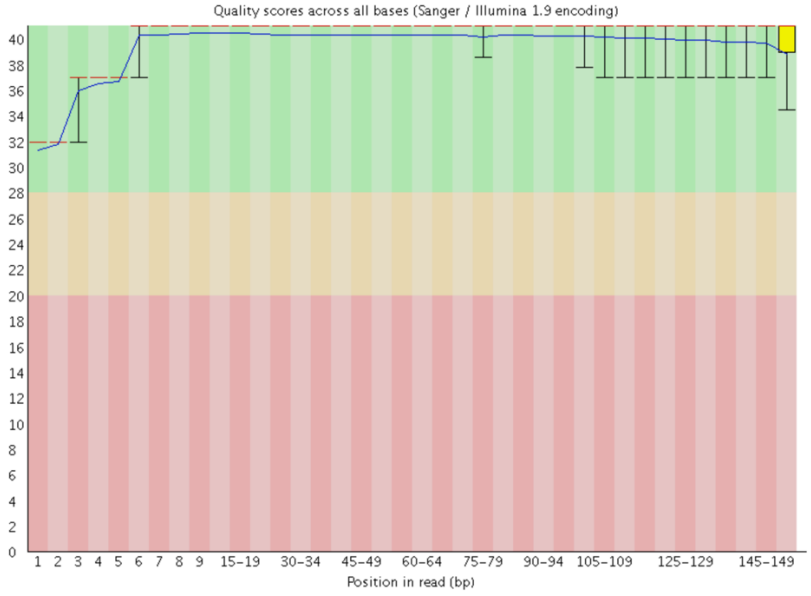


Figure 5.8 The representative image from western blot assay of the immunoprecipitated FLAG-tagged CPVIB-1 protein.

(A) the western blot image using anti-FLAG primary antibody, (B) the western blot image using anti-Ub antibody.
Lane M, 15 µl of Bio-Rad protein ladder,
lane 1, the immunoprecipitated FLAG-tagged CPVIB-1 protein from 800 µg total proteins of the positive FLAG-tagged CPVIB-1 strain (positive control),
lane 2, the immunoprecipitated FLAG-tagged CPVIB-1 protein from 800 µg total proteins from the positive FLAG-tagged CPVIB-1 strain treated with 10 ng/ml rapamycin
lane 3, the immunoprecipitated FLAG-tagged CPVIB-1 protein from 800 µg total proteins from EP155 strain (negative control),
lane 4, the immunoprecipitated FLAG-tagged CPVIB-1 protein from 800 µg total proteins from Δ*cpvib-1* strain (negative control).



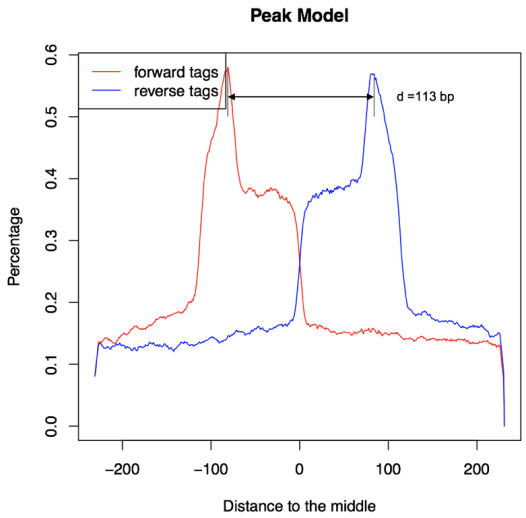
(A)



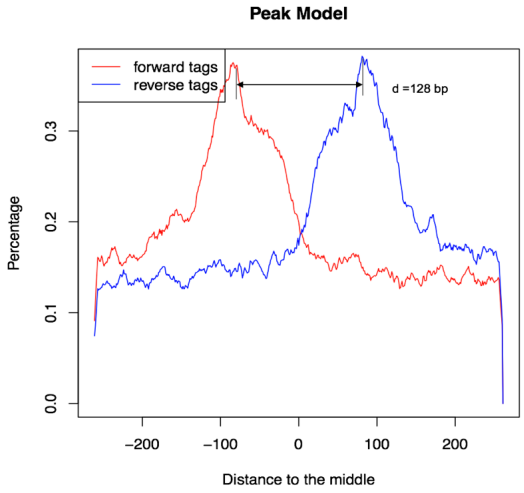
(B)

Figure 5.9 Per base sequence quality plot of sample 1 from FastQC.

X axis is the position in reads (150 bp for each read), Y axis is the sequencing quality score (Above 30 are considered high quality) (A) The plot was from the original reads of sample 1. (B) The plot was from the reads of sample 1 after trimming process.



(A)



(B)

Figure 5.10 MACS2 models for peaks in FLAG-tagged-CPVIB-1 samples and the FLAG-tagged-CPVIB-1-Rapamycin samples.

X axis is the distance to the middle point of the peaks, and Y axis is the distribution percentage of tags.
(A) The 5' strand-separated tags from the FLAG-tagged-CPVIB-1 sample were aligned by the center of their forward and reverse peaks. (B) The 5' strand-separated tags from the FLAG-tagged-CPVIB-1-Rapamycin sample were aligned by the center of their forward and reverse peaks.

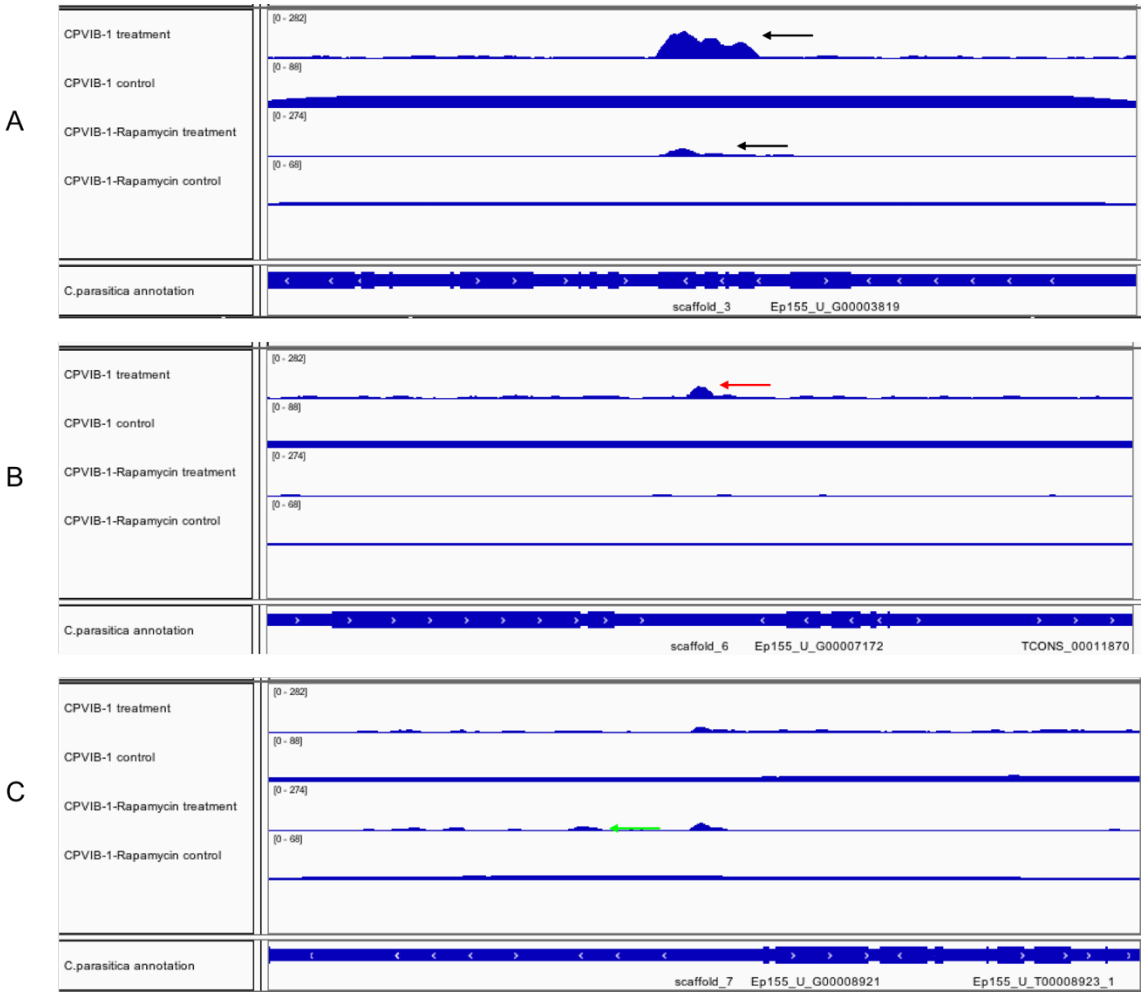


Figure 5.11 Three examples of peaks identified in the promoter regions of annotated genes.

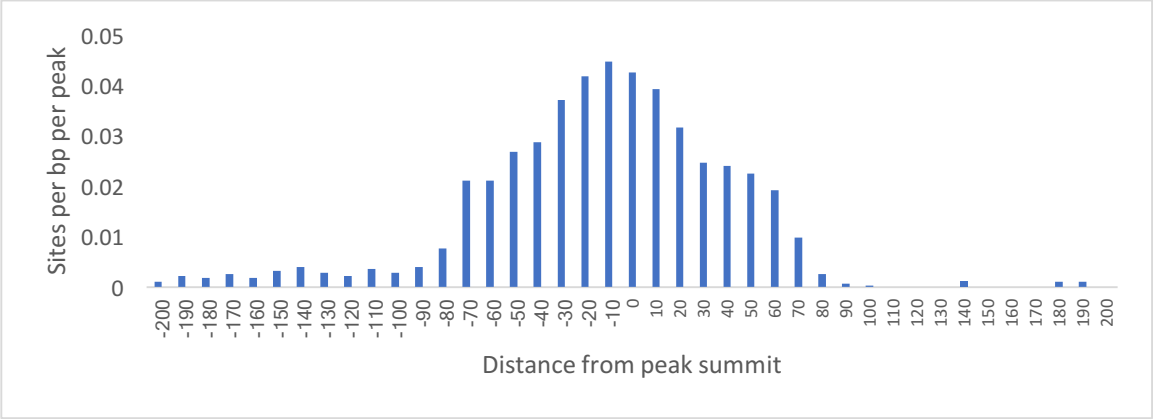
(A) Two peaks were identified in the promoter of the same gene from both FLAG-tagged CPVIB-1 samples and its Rapamycin treatment samples (highlighted with black arrows). (B) The peak was identified in the promoter of the gene from only FLAG-tagged CPVIB-1 samples (highlighted with red arrows). (C) The peak was identified in the promoter of the gene from only FLAG-tagged CPVIB-1-Rapamycin samples (highlighted with green arrows).

Information for motif1

Reverse Opposite:

p-value:	1e-20
log p-value:	-4.760e+01
Information Content per bp:	1.530
Number of Target Sequences with motif	68.0
Percentage of Target Sequences with motif	24.73%
Number of Background Sequences with motif	2486.6
Percentage of Background Sequences with motif	6.63%

(A)



(B)

Figure 5.12 (Continued)

Rank	Motif	Name
1		BPC6(BBRBPC)/col-BPC6-DAP-Seq(GSE60143)/Homer
2		FRS9(ND)/col-FRS9-DAP-Seq(GSE60143)/Homer
3		BPC1(BBRBPC)/colamp-BPC1-DAP-Seq(GSE60143)/Homer
4		VRN1(ABI3VP1)/col-VRN1-DAP-Seq(GSE60143)/Homer
5		GAGA-repeat/Arabidopsis-Promoters/Homer

(C)

Figure 5.12 The predicted recognition sequences of FLAG-tagged CPVIB-1 samples.

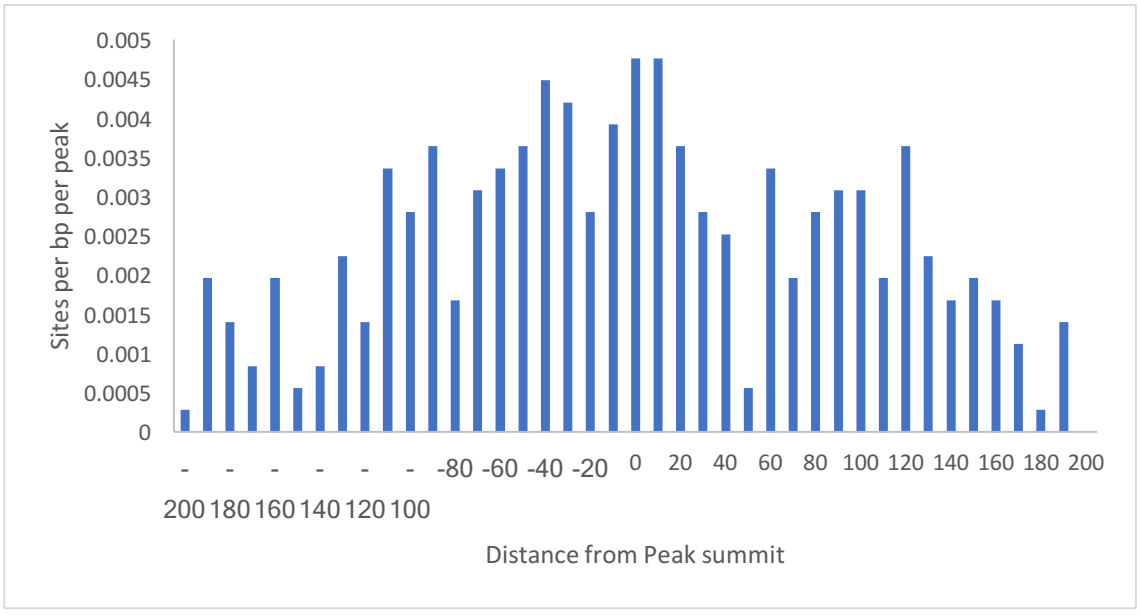
(A) The GAGAGAGA recognition sequence of FLAG-tagged CPVIB-1 protein. (B) The distance between the center of predicted recognition sequences and the peaks summit. (C) The top rank HOMER known genes with similar recognition sequences of FLAG-tagged CPVIB-1 protein.

Information for 1-CTTCTTYC (Motif 1)

Reverse Opposite:

p-value:	1e-7
log p-value:	-1.765e+01
Information Content per bp:	1.652
Number of Target Sequences with motif	116.0
Percentage of Target Sequences with motif	32.49%
Number of Background Sequences with motif	7826.0
Percentage of Background Sequences with motif	20.04%

(A)



(B)

Figure 5.13 (Continued)

Rank	Motif	Name
1		KLF10(Zf)/HEK293-KLF10.GFP-ChIP-Seq(GSE58341)/Homer
2		FRS9(ND)/col-FRS9-DAP-Seq(GSE60143)/Homer
3		BPC6(BBRBPC)/col-BPC6-DAP-Seq(GSE60143)/Homer
4		Trl(Zf)/S2-GAGAfactor-ChIP-Seq(GSE40646)/Homer
5		GAGA-repeat/Arabidopsis-Promoters/Homer

(C)

Figure 5.13 The predicted recognition sequences of FLAG-tagged CPVIB-1-Rapamycin samples.

(A) The similar (GA)n recognition sequences of FLAG-tagged CPVIB-1-Rapmaycin samples. (B) The distance between the center of predicted recognition sequence and the peaks summit. (C) The top rank HOMER known genes with similar recognition sequences of FLAG-tagged CPVIB-1 protein with the treatment of rapamycin.

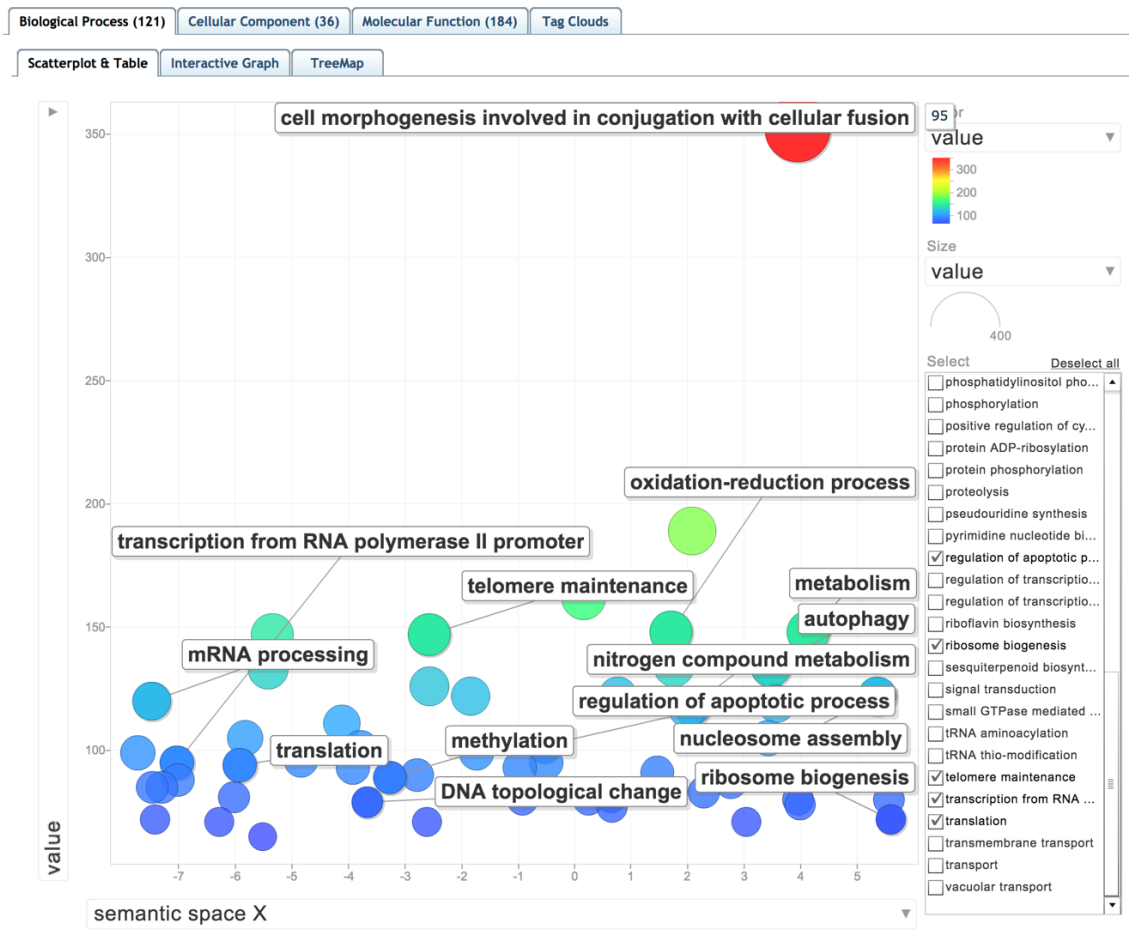


Figure 5.14 The "Scatterplot & Table" view of REVIGO showing the GO clusters of the direct targeted genes by FLAG-tagged CPVIB-1 protein.

The y axis represents the peak score of each GO term.
The x axis represents the semantic scale of the GO terms.
The bubble color indicates the log2 fold change value from the positive (red, up-regulated) to negative (blue, green, and yellow down-regulated). The size of the bubble indicates the frequency of the GO term in the dataset.

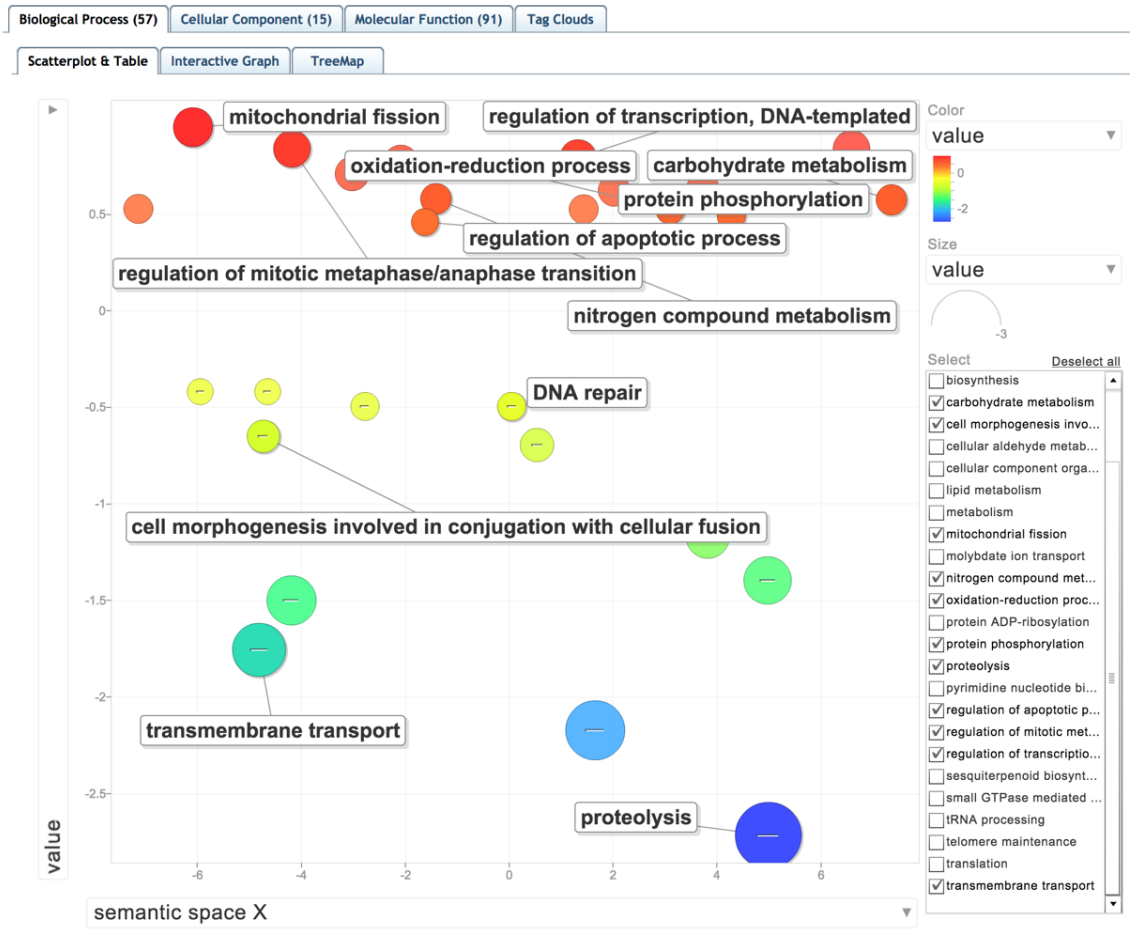


Figure 5.15 The "Scatterplot & Table" view of REVIGO showing the GO clusters of the direct targeted genes by FLAG-tagged CPVIB-1 protein and significantly altered in transcriptional level in the Δ*cpvib-1* mutant strain.

The y axis represents the log2 fold change of each GO term.
The x axis represents the semantic scale of the GO terms.
The bubble color indicates the log2 fold change value from the positive (red, up-regulated) to negative (blue, green, and yellow down-regulated).
The size of the bubble indicates the frequency of the GO term in the dataset.

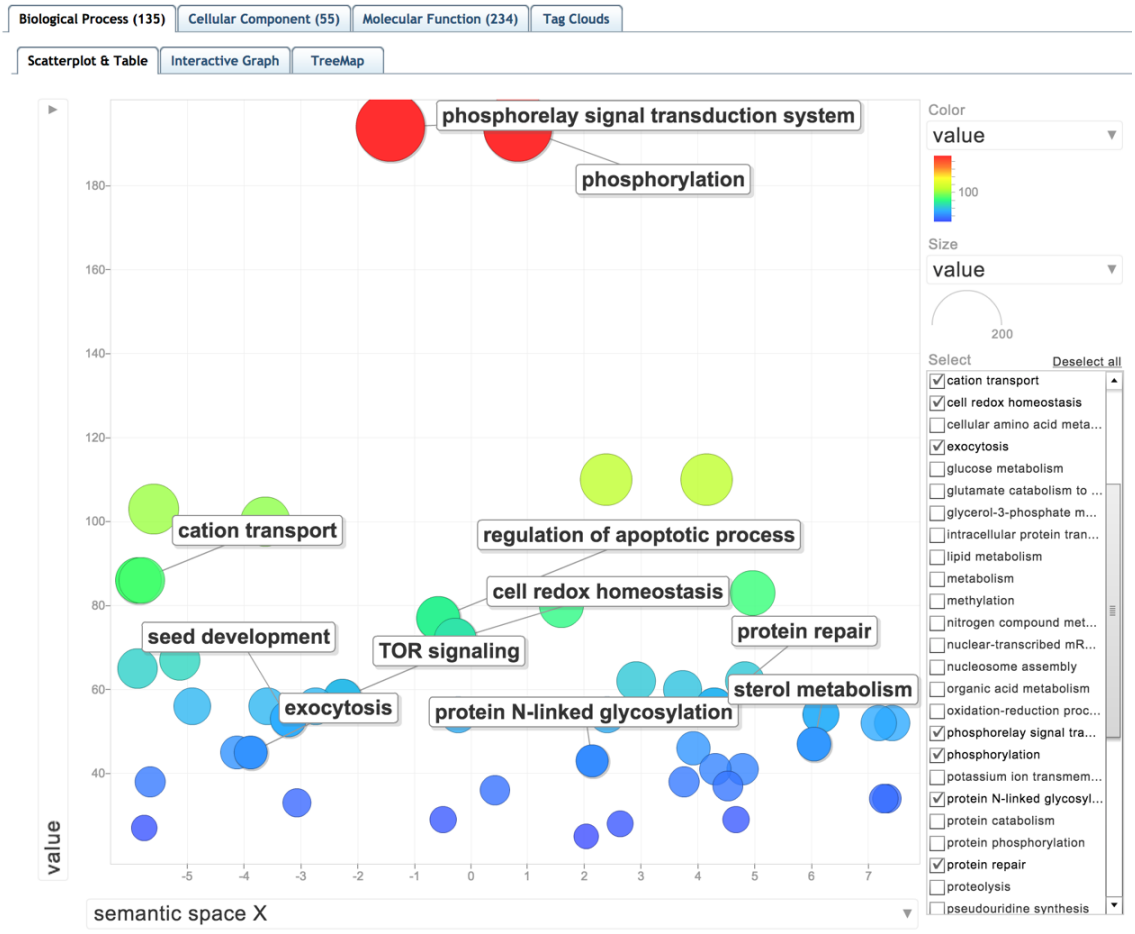


Figure 5.16 The "Scatterplot & Table" view of REVIGO showing the GO clusters of the direct targeted genes by FLAG-tagged CPVIB-1 protein with rapamycin treatment.

The y axis represents the peak score of each GO term.
The x axis represents the semantic scale of the GO terms.
The bubble color indicates the log2 fold change value from the positive (red, up-regulated) to negative (blue, green, and yellow down-regulated).
The size of the bubble indicates the frequency of the GO term in the dataset.

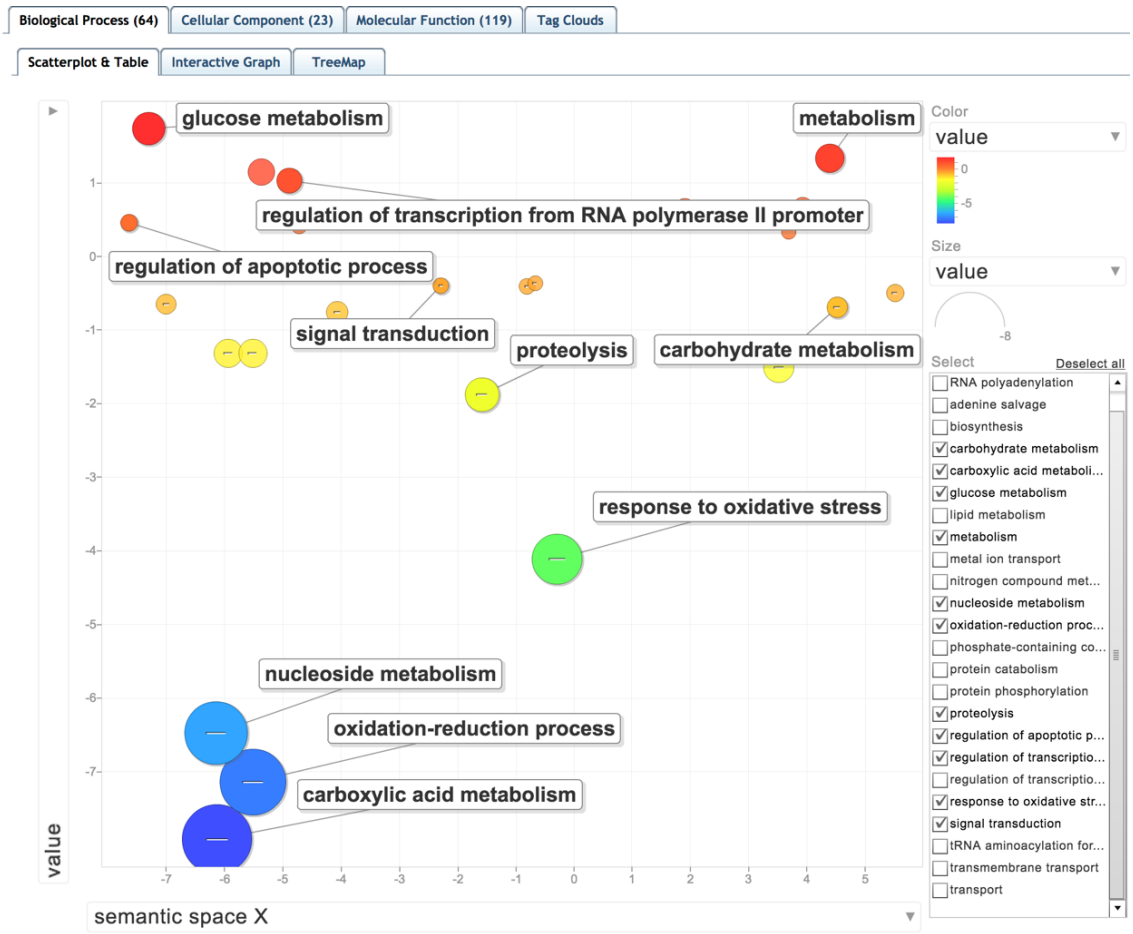


Figure 5.17 The "Scatterplot & Table" view of REVIGO showing the GO clusters of the direct targeted genes by FLAG-tagged CPVIB-1 protein with rapamycin treatment and significantly altered in transcriptional level in the Δ*cpvib-1* mutant strain.

The y axis represents the log2 fold change of each GO term.
The x axis represents the semantic scale of the GO terms.
The bubble color indicates the log2 fold change value from the positive (red, up-regulated) to negative (blue, green, and yellow down-regulated).
The size of the bubble indicates the frequency of the GO term in the dataset.

CHAPTER VI

DISCUSSION

Significance of a re-annotated *C. parasitica* genome

In 2002, Ouzounis defined the term, re-annotation, stating the trend and requirement for genome re-annotation to discover more gene and protein functions, test and refine the mistakenly predicted genes, and perform the comparison of existing and newly developed methods [148]. Next generation sequencing (NGS) technologies create great opportunities for various genomics, transcriptomics, and proteomics projects of exotic non-model organisms [80; 149]. Meanwhile, the drive for re-annotation of the prior existing genomes arises with the up-to-date evidence, such as transcriptomic and proteomic datasets as well as the newly developed bioinformatics tools. [150]. As the most significant factor for judging the various versions of annotation, quality assessment always represents a great challenge and raises a serious issue since then [148]. For example, the genome of *C. parasitica* was sequenced, annotated, and released in 2009 using the technology and tools available at that time, but did not contain any quality assessment step.

In this study, with deep transcriptome sequencing and updated protein evidence, re-annotation of *C. parasitica* genome is more accurate and informative compared to the prior version. Furthermore, since MAKER2 is the first bioinformatics tool to provide a quality metrics system, the quality assessment provided a means to highlight any

problematic predicted gene models for further manual curation and also gave a measure for the comparison of two annotation versions.

This significantly improved genome annotation (2017-version) not only serves the purpose to provide accurate and informative results for characterizing the roles of CPVIB-1 using RNA-Seq and ChIP-Seq projects, but also provides a valuable resource for researchers who are interested in both comparative and functional studies of *C. parasitica*. The integrated annotation analysis by applying the MAKER2-two-pass pipeline has facilitated the improvement of the genome annotation, and this approach can be applied to other biological systems.

Development and application of PEPA, a prior genome annotation evaluation pipeline

In 2009, Bakke reported a comparison of three automated genome annotations, the Joint Genome Institute (JGI), the National Microbial Pathogen Data Resource (NMPDR) and the J. Craig Venter Institute (JCVI) [151]. Later in 2013, a semi-automated genome annotation comparison scheme was developed to perform the functional comparison only [75]. In 2011, MAKER2 was developed to attach a quality assessment system to the automated genome annotation, which providing a means to compare the annotations in quality aspect besides the structural and functional features [48].

PEPA is a pipeline designed for the evaluation of any pre-existing genome annotation with no quality metric system. It is a simple pipeline that was developed to comprehensively estimate the accuracy of each prior predicted gene model, to provide updated transcripts and protein evidence using the MAKER2 legacy annotation program,

and enrich the gene models with a quality metrics system and internal domains (InterPro ID). It also contained a R script to perform a comparison of the prediction quality of all gene models from the prior annotation and the newer version. Finally, it included a python script that was developed to sort each individual gene model from the prior genome annotation into four categories (Match, Similar, Different, Noexist) based on their discrepancies with the newer version.

The PEPA evaluation pipeline has a significant benefit for the future work in *C. parasitica* by providing reliable insights of each predicted gene model that researchers are interested in with the new quality metrics system and sorting system. It also has profound impact to the genome projects of exotic non-model organisms, in which the structural and functional accuracy of a predicted gene model is extremely important for characterizing it by developing the knock-out constructs. To the best of our knowledge, a comprehensive pipeline designed to evaluate a prior genome annotation, regardless of the format against a newer version, has not been presented yet to illustrate the structural and functional differences of each gene model and its accuracy.

Structure of CPVIB-1

CPVIB-1 is a NDT80/PhoG-like transcription factor

CPVIB-1 is predicted to be a putative p53-like transcription factor from the InterPro Protein sequence analysis and classification website (Figure 6.1) [152]. The NDT80-DNA binding domain (PRU00850) is found in proteins with size ranging from 185-382 amino acids that have transcription factor activity and many are thought to response to celluar nutritional status (https://prosite.expasy.org/rule/PRU00850). NDT80-like genes are found only in the Amorphea (unikont) lineage, which includes animals,

fungi and Amoebozoa [31]. The number of these genes present in the genome varies in the fungi, ranging from zero to six in the Ascomycota class [153]. Before this study, the analysis of NDT80-like genes has only been completed in three fungal species, the haploid ascomycetes *S. cerevisiae*, *Aspergillus nidulans*, and *N. crassa*, which contain one, two and three NDT80-like genes, respectively [30; 154]. In *C. parasitica*, there are three NDT80-like genes identified in the 2017-version annotation: Ep155_U_G00001434, similar to transcription factor vib-1; Ep155_U_G00009104, similar to transcription factor pacG; and, Ep155_U_G00004090, similar to meiosis-specific transcription factor NDT80. All of them were predicted to be transcription factors and orthologs of VIB-1 gene in *N. crassa*, pacG gene in *A. nidulans,* and NDT80 gene in *S. cerevisiae*, respectively [32; 154-155].

The NDT80-DNA binding domain in yeast is comprised of a β-sandwich core with seven β-strands and three short α-helices and shows high structural conservation with similar domains from *N. crassa*, *D. melanogaster*, and humans, and is absolutely conserved in the positions that are identified to interact with DNA [156]. The SWISS-MODEL was used to predict the quaternary structure of CPVIB-1 and the same β-sandwich core was observed with an additional six β-strands and short α-helices (https://swissmodel.expasy.org/) (Figure 6.2). In this study, CPVIB-1 is the first NDT80-DNA binding transcription factor in *C. parasitica* to be characterized by its structure and functions.

Ubiquitin decoration in CPVIB-1

Transcription factors comprise a relatively small portion of genome, approximately 7% in humans and 1.1% in *C. parasitica* (2017-version annotation) and function as the gatekeepers of cellular functions, integrating signal transformation into the activation/repression of gene expression that reconfigures the cell physiology in real time [157]. Post-translational modifications (PTMs) regulate the function, abundance, and activity of the transcription factors in every aspect [157]. Among all the PTMs, the more prominently studied are phosphorylation, methylation, glycosylation, acetylation, sumoylation, and ubiquitination [129; 157-160]. In all the well-studied PTMs, the ubiquitination modified transcription factors are found to be the least prevalent in humans[157]. Ubiquitin is a small protein that is covalently linked to the lysine side chain of substrates by the cognate E3 ligase. This may occur at single or multiple lysine residues [161]. Generally, poly-ubiquitination of transcription factors marks them for degradation through the proteasomal system, which is called the ubiquitin-dependent proteolysis pathway [129; 159].

In this study, the ubiquitin modification of CPVIB-1 was predicted using UbSite online tool (http://csb.cse.yzu.edu.tw/UbiSite/), where the protein sequence of CPVIB-1 was predicted to have eight ubiquitination sites (Ubi-sites) with relatively high confidence scores (Figure 6.3). Consistent with the above identification, CPVIB-1 was found to be covalently attached with at least one ubiquitin proteins from western blot assay in one or multiple sites, and its ubiquitinated-version comprised the major portion of total cellular CPVIB-1. It is most likely that the non-ubiquitinated version is the one with DNA binding activity in genome. Based on the difficulties experienced in extracting

sufficient DNA bound to CPVIB-1 protein compared to other ChIP-Seq projects, it can be speculated that the poly-ubiquitination of CPVIB-1 leads to the degradation through the classic proteolysis pathway, which regulates the level of active CPVIB-1 protein in cells [99; 109].

Functions of CPVIB-1

NDT80-DNA binding transcription factor CPVIB-1

In budding yeast (*S. cerevisiae*), the NDT80-like protein is required for the completion of meiosis [154]. In the pathogenic *Candida albicans*, the CaNDT80-like protein is required for antifungal drug resistance, hyphal growth, biofilm formation, and virulence [162-164]. In the filamentous fungus *A. nidulans*, NDT80-like XprG is a regulator of a large number of genes involved in carbon starvation, extracellular protease, mycotoxin and penicillin production [31; 153; 165]. In *N. crassa*, VIB-1 is required for expression of genes that are involved in vegetative incompatibility-induced programmed cell death and is also a regulator for the genes involved in nutrient starvation [32; 34; 79]. In *C. parasitica* the ortholog of the above NDT80-like transcription factors, CPVIB-1, is a regulator of vegetative incompatibility programmed cell death, hyphal growth, sporulation, and virulence.

This study is the first to attempt characterization of the targets of the NDT80-like transcription factor CPVIB-1. By using large-scale transcriptome comparison analysis, the regulating mechanisms behind the above functions can be understood. For example, the role of CPVIB-1 in carbon metabolism CPVIB-1 has been identified to regulate at least five various but directly related KEGG carbon pathways that repress glucose utilization. Besides, the accumulation of CPVIB-1 is increased with the glucose

starvation treatment using western blot assay that is congruent with the above transcriptome analysis results. Meanwhile, CPVIB-1 has also been determined to regulate various genes related to the pathogenesis. Perhaps the regulation of carbon metabolism-related genes by CPVIB-1 is an important response during the plant host infection process, which could explain the reduced virulence of the *Cpvib-1* deletion strain (Figure 6.4). However, we also found that CPVIB-1 is a universal regulator that is essential for RNA process, cellular biosynthesis processes, responses to stress, and hundreds of other genes with diverse functions (Figure 6.4). With the studies from the NDT80-like transcription factors in the above model fungal species, transcriptomics alone failed to provide insights to further explain the targets of CPVIB-1. Therefore, we performed the ChIP-Seq to explore the direct targeted genes of CPVIB-1.

GAGA factor CPVIB-1

In *Drosophila*, GAGA factor (DmGAF) has an extraordinarily diverse set of functions that include the activation and silencing of gene expression, nucleosome organization, remodeling chromosome architecture and mitosis by directly interacting with a small subset of target genes and indirectly interacting with many others to serve varied functions [96]. In vertebrates, all GAGA factors (vGAF) contains N-terminal BTB/POZ domains, four C-terminal zinc finger domains for the $(GA)^n$ specific DNA binding, and multiple PTM sites for phosphorylation, acetylation, and ubiquitination. Similar to the function of DmGAF in *Drosophila*, vGAF play roles in gene activation, repression, enhancer blocking, and cell development and differentiation [166].

In this study, using the ChIP-Seq method, CPVIB-1 has been identified as the first GAGA factor in the fungal kingdom. CPVIB-1 was found to bind to sequences associated

with 264 genes that function in remarkably diverse range of regulatory contexts, including activation/repression of transcription, mitosis maintenance, cell development, autophagy, cell-cell communication, virulence, nucleosome assembly, and others. Furthermore, CPVIB-1 was found to bind to sequences thought to control a DNA helicase for mitosis, telomere maintenance and the decomposition of the chromatin complex as well as the activators for the transcription of RNA polymerase II promoter for the promotion of the transcription process. Like the GAGA factors in Drosophila and human, CPVIB-1 is a multifaceted transcription factor with diverse roles in gene activation and repression, maintenance of mitosis and cell development (Figure 6.5) [166].

CPVIB-1 functions in the TOR signaling pathway

Our interest in the role and function of CPVIB-1 stemmed from the observation that this protein is required for the response to certain vegetative incompatibility responses. In *P. anserina*, the *idi* gene, which is induced during programmed cell death and triggered by vegetative incompatibility, was also found to be induced upon nitrogen and carbon starvation [105]. In the same study, a well-known pathway response to nutrient signals, the TOR kinase pathway, was linked to the vegetative incompatibility and rapamycin was demonstrated to be capable of inducing the expression of *idi* gene to mimic the vegetative incompatibility [105].

In this study, rapamycin and the carbon starvation was found to induce the accumulation of CPVIB-1 protein. Moreover, CPVIB-1 was found to positively correlated with the action of rapamycin to inhibit the fungal cell growth. To explore the mechanisms behind the link of vegetative incompatibility and TOR signaling pathway,

the ChIP-Seq method was applied to identify the recognition sequence bound by the CPVIB-1 protein. The recognition sequence of CPVIB-1 upon rapamycin treatment was found to be slightly shifted to a flexible $(GA)^n$ motif and, therefore, the annotated genes associated with this altered recognition sequence were shifted as well. In particular, CPSTE20, a TOR2 complex subunit, and CPRGA1, a Rho-type GTPase-activating protein, were found to be in this adjusted set of CPVIB-1 protein targets.

In yeast Ste20p, also known as the ortholog of Avo3p, is the largest subunit of TORC2 complex and the C-terminal part of Ste20p, is located close enough to the rapamycin binding domain of TORC2 to protect it from the binding of rapamycin in yeast [167-168]. Rga1p plays a crucial role in activating multiple GTPase genes, major components in the MAPK cascade for the activation of the TOR2 signaling pathway [169-170] . Both of these genes code for critical proteins involved in the TOR signaling pathway and the data presented here indicates they are regulated directly by CPVIB-1 upon the treatment of rapamycin.

A new interpretation of the TOR signaling pathway that includes CPVIB-1

In yeast, the macrolide antifungal chemical rapamycin was found to bind to peptidyl-prolyl cis-trans isomerase, also as known as Fpr1p. [101; 171-173]. Subsequently, from the mutants that conferred resistance to rapamycin, the target of rapamycin was identified [171; 174]. There were two novel proteins, Tor1p (target of rapamycin 1) and Tor2p (target of rapamycin 2), that were found to physically interact with the Fpr1-rapamycin complex in the FRB domain (Fpr1-rapamycin binding domain) [175-178].

The domain structure and amino acid sequences of Tor1p and Tor2p are evolutionarily conserved [175]. In yeast, both of them contain the HEAT repeats, FAT, FRB, kinase, FIT and FATC (Figure 6.6). Although the primary structure of the Tor1p and Tor2p are highly similar to each other, their cellular functions are distinct [175; 179]. The kinase activity of TOR complex 1 (TORC1) is found to be essential for cell growth by promoting anabolism, such as protein biosynthesis, lipid biosynthesis and nucleotide biosynthesis for the need of new cells by phosphorylating proteins, promoting mRNA translation initiation process, promote ribosome biosynthesis [102; 113]. TORC1 also represses catabolism, such as autophagy and ubiquitin-proteasome system [102]. While TORC1 regulates cell growth and metabolism, the kinase activity of TOR complex 2 (TORC2) controls proliferation and survival process by phosphorylating several downstream protein kinases and regulating the glucose utilization in humans by responding to insulin signal [102; 113].

This functional distinction between them is due to the differences in the composition of the TOR complexes. TORC1 is composed with either Tor1p or Tor2p as the back-bone protein, and Kog1p, Tco89p, and Lst8p are subunits (Figure 6.5) [180-182]. In TORC2, there are Tor2p is always the TOR protein and is assembled with the subunits of Avo1p, Avo2p, Avo3p, Bit61p, and Lst8p (Figure 6.6) [183-186]. Avo3p is found to be an essential scaffold protein to assemble TORC2 and interact with FAT and kinase domains of Tor2p within TORC2 to prevent the accessibility from the FKBP12-rapamycin complex [186]. Because of this, TORC1 is sensitive to rapamycin treatment, since it lacks the Avo3p component, but TORC2 was found to be resistant to rapamycin treatment [178].

In *C. parasitica*, CPVIB-1 was associated with the promoter region of *cpste20*, encoding a ortholog protein of Avo3p following the treatment with rapamycin [168]. The binding recognition motif of CPVIB-1 in this promoter region is GAGAAGAGC following the treatment with rapamycin, which is slightly shifted from GAGAGAGA under untreated conditions. Rapamycin is known to target TORC1 to inactivate its function. We have shown that CPVIB-1 protein accumulated to higher levels following rapamycin treatment or glucose starvation. Our data from RNA-Seq showed that in the Δ*cpvib-1* strain glucose utilization is significantly and consistently upregulated. Therefore, we propose that a function of CPVIB-1 is to repress this pathway, opposite to the role of TORC2. Similarly, CPVIB-1 appears to be required for the repression of protein, lipid, and steroid biosynthesis, and activation of autophagy and proteolysis, which is opposite to TORC1. Furthermore, our data also shows by ChIP-Seq that CPVIB-1 can bind to the promoter regions of *cpste20*, which encoding protein to protect TORC2 from rapamycin action, but the regulation interplay between TORC1 and TORC2 was observed but not understood. Based on the results of this study, we proposed a revised model of TOR signaling that places CPVIB-1 between the TORC1 and TORC2. To be responsible for inducing the cell death autophagy in vegetative incompatibility, CPVIB-1 is repressed by TORC1 under normal conditions. However, during the nutrient starvation or rapamycin treatment, TORC1 is initially inactivated, which led to the increased accumulation of CPVIB-1. In turn, CPVIB-1 bound the promoter of *cpste20* repressing its transcription and, presumably its protein abundance, thereby resulting in less protection of TORC2 from rapamycin. Ultimately, rapamycin is then able to inactivate

both TORC2 and TORC1 resulting in more profound effects of inhibition of cell growth, cell survival and proliferation (Figure 6.7).

Future directions

Because our predicted model of the function of CPVIB-1 in TOR signaling pathway superseded our earlier transcriptome work, we do not have transcriptome profiling data of *C. parasitica* following treatment with rapamycin. Therefore, an immediate first strategy would be to perform transcription level comparison of *cpste20* between *C. parasitica* EP155 strain and its isogenic *Δcpvib-1* strain under the same conditions of rapamycin treatment as the ChIP-Seq experimental design. Also, it is necessary to perform the comparison analysis against to the EP155 strain and its isogenic *Δcpvib-1* strain without rapamycin treatment. If rapamycin targets TORC1 leading to the induction of CPVIB-1, leading to the repression of *cpste20*, rendering TORC2 vulnerable to rapamycin, we can predict that the *cpste20* will be repressed in transcript level in EP155 strain in rapamycin treatment with the presence of CPVIB-1 compared to in *Δcpvib-1* strain upon the same condition of rapamycin treatment without the presence of CPVIB-1 as well as in EP155 strain without the rapamycin treatment.

A second option would be to test the potential for CPSTE20 protein detection using the commercial anti-STE20 antibody, subdomain VI from yeast (Sigma-Aldrich, USA). If this was cross-species functional as a detection tool, we could test the prediction that the accumulation of CPSTE20 protein will be decreased in EP155 strain following rapamycin treatment in the presence of CPVIB-1 compared to the *Δcpvib-1* strain or in the EP155 strain without rapamycin treatment.

The TOR signaling system is known as one of the signal network systems that has been evolutionarily conserved among eukaryotes [103; 187; 188]. In higher eukaryotes, such as in humans, mammalian TORC1 (mTORC1) plays a central role in regulating all of the downstream factors and therefore controls the balance between anabolism and catabolism in response to the environmental conditions [113]. Dysfunctions in the TOR signaling network can closely correlate with pathological conditions including diabetes, cancer, obesity, and neurodegeneration [187]. However, for the TORC2, because the knockout of TOR2 gene in yeast is lethal and TORC2 is resistant to rapamycin, the studies of upstream and downstream factors of TORC2 signaling haven't been extensively investigated [113; 189]. Here we present a potential inhibition strategy for TORC2, rapamycin with the CPVIB-1, in *C. parasitica*. With this strategy, an immediate third idea would be to perform the transcriptome sequencing of *C. parasitica* EP155 strain and its isogenic Δcpvib-1 strain under the same conditions of rapamycin treatment as the ChIP-Seq experimental design. With the transcriptome data available, it would be possible to perform the comparison analysis against to the transcriptome EP155 strain and its isogenic *Δcpvib-1* strain as in Chapter III. The transcriptome comparison analysis can be used to reveal the downstream factors of TORC2 signaling pathway with the above inhibition strategy. Meanwhile, the genes downstream of TORC1 are responsible for regulating CPVIB-1 can also be revealed using the same strategy. Furthermore, the transcriptome comparison analysis of the EP155 strain with and without rapamycin treatment can be used to understand the mechanism of TORC1 slightly shifting the recognition sequence of CPVIB-1 as well.

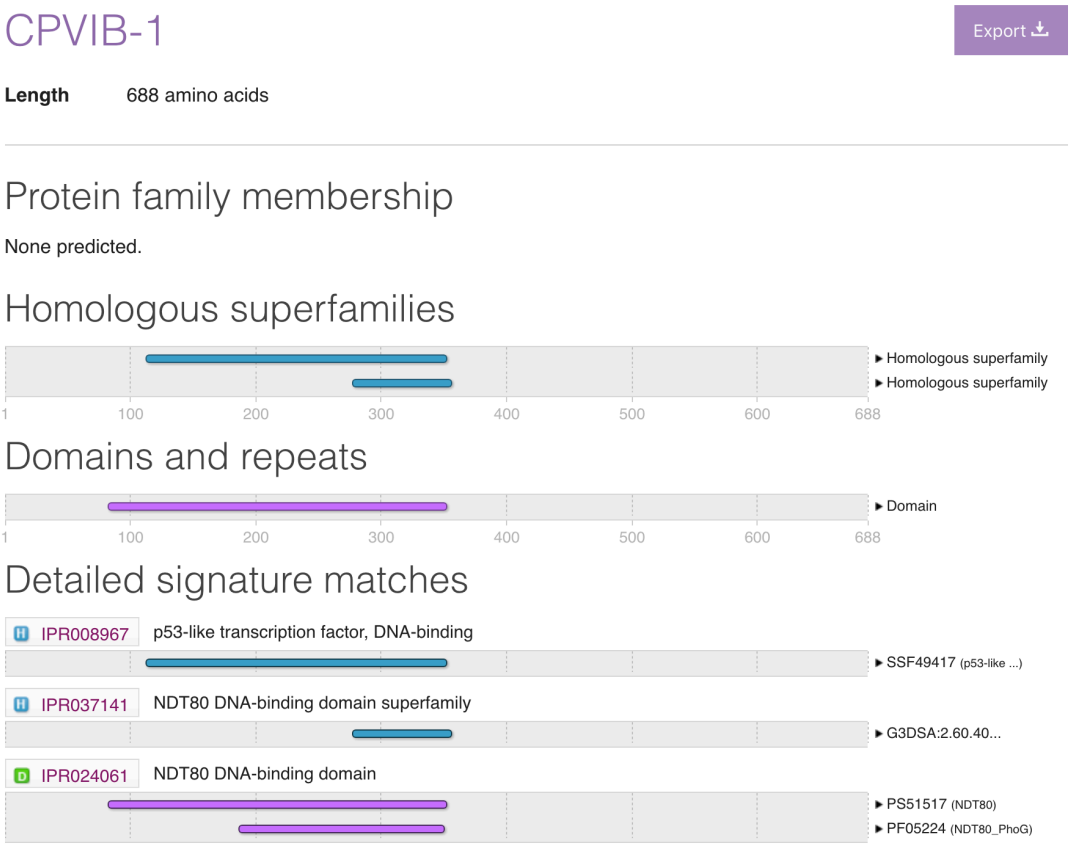


Figure 6.1 InterPro Protein sequence analysis & classification prediction results of CPVIB-1 protein.

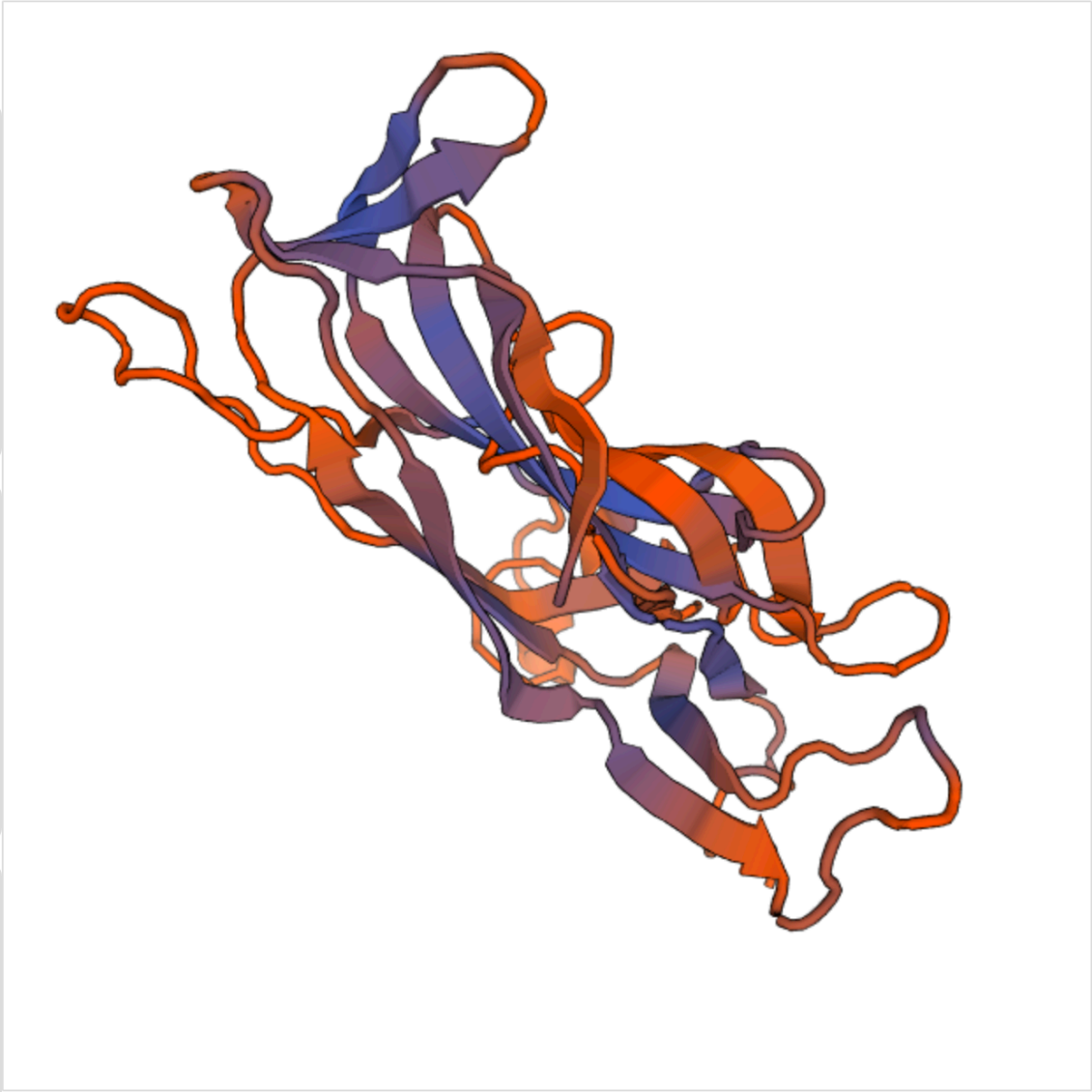


Figure 6.2 The predicted quaternary structure of CPVIB-1 from SWISS-MODEL.

Download Result	Input Information
Threshold	High
Input ID	CPVIB-1
Input Sequence	MCTATYAVTMAELKEPASQHNLWYGSPVQMSTSRPSHAADAGLSS...

Predict Result

Order	Protein Name	Locations	Score (Confidence)	Ubiquitination Sites	Substrate Motifs
1	CPVIB-1	129	0.506789 (High)	LRDDGT K LGVGHD	
2	CPVIB-1	147	0.649383 (High)	FRKVPD K LTILDG	
3	CPVIB-1	328	0.515591 (High)	TLANGA K AVLSEA	
4	CPVIB-1	353	0.565957 (High)	RNFQAR K EIPLLG	
5	CPVIB-1	384	0.730516 (High)	AGPLSV K TQDSKG	
6	CPVIB-1	389	0.516561 (High)	VKTQDS K GRPMNI	
7	CPVIB-1	452	0.726404 (High)	ASDPYQ K LPLSGT	
8	CPVIB-1	664	0.526475 (High)	KTSAGV K NDPHAP	

Figure 6.3 The Ubiquitination sites identified for CPVIB-1 protein sequences in the UbiSite web server.

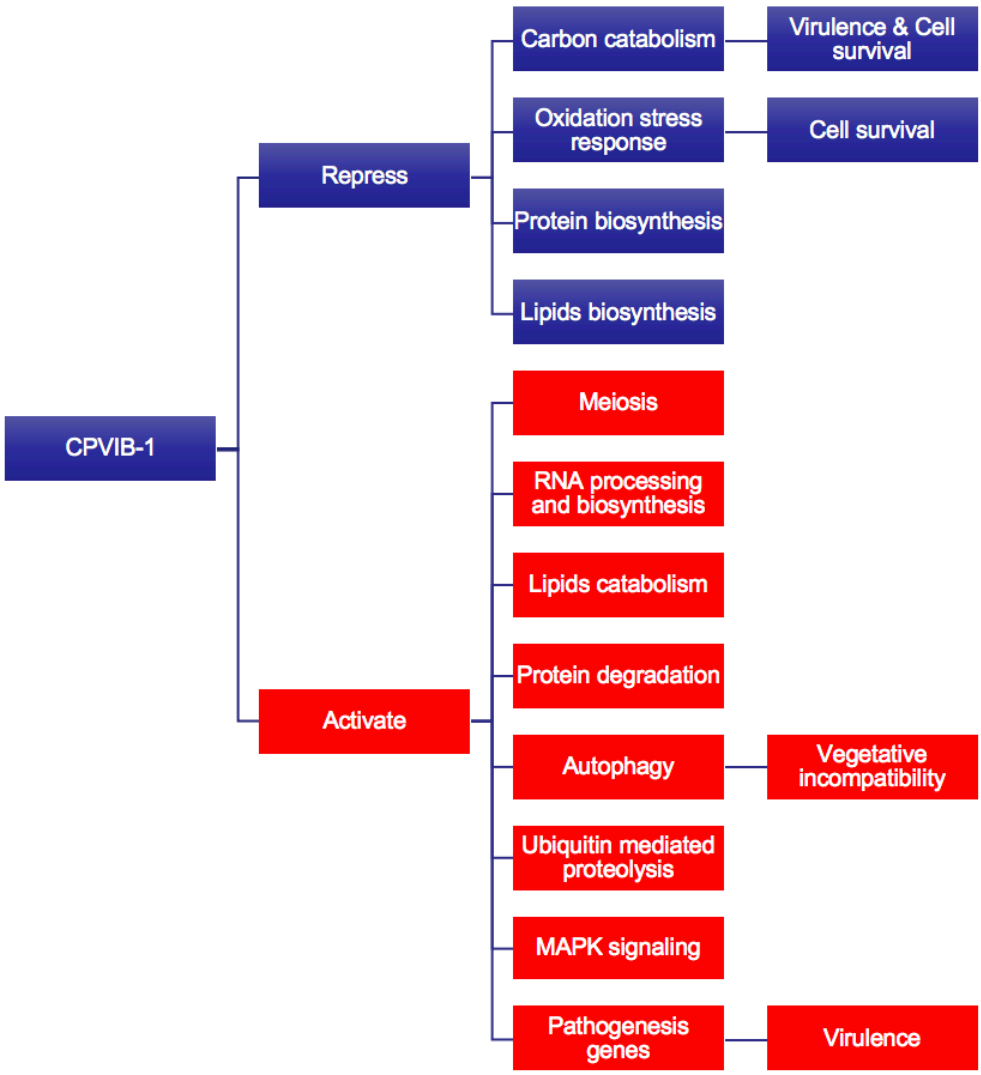


Figure 6.4 The biological processes regulated by CPVIB-1 from the transcriptome comparison analysis.

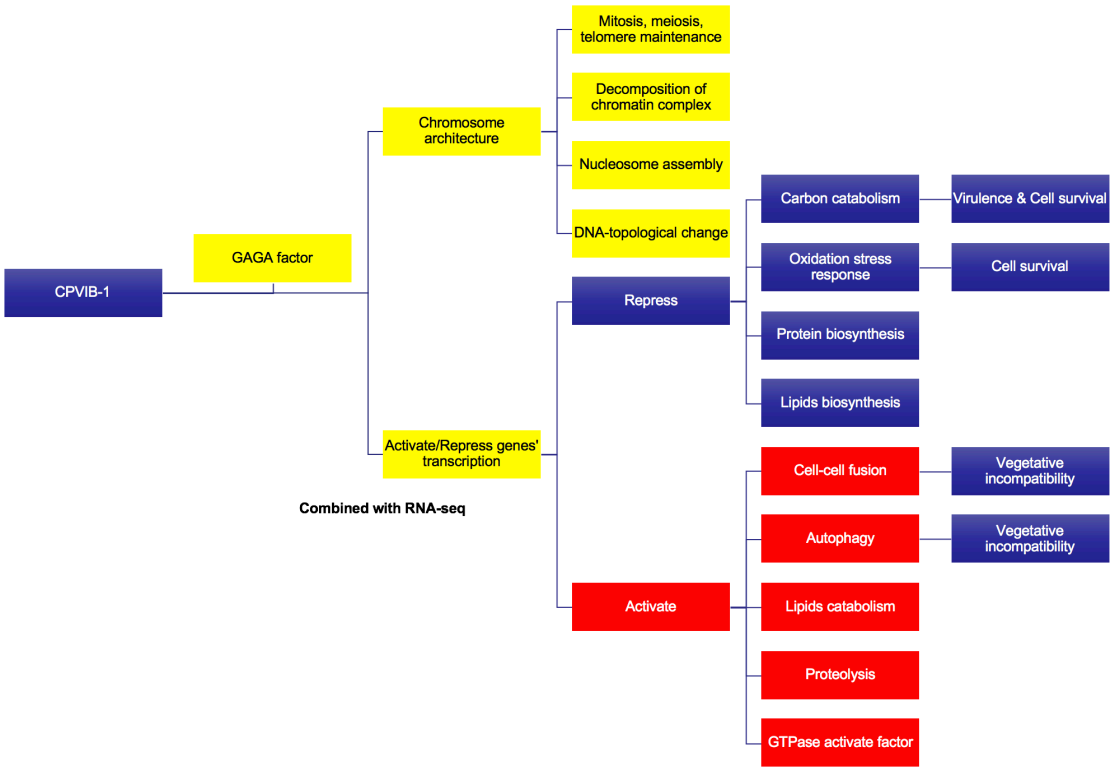


Figure 6.5 The biological processes regulated by CPVIB-1 from the ChIP-Seq analysis combined with transcriptome comparison analysis.

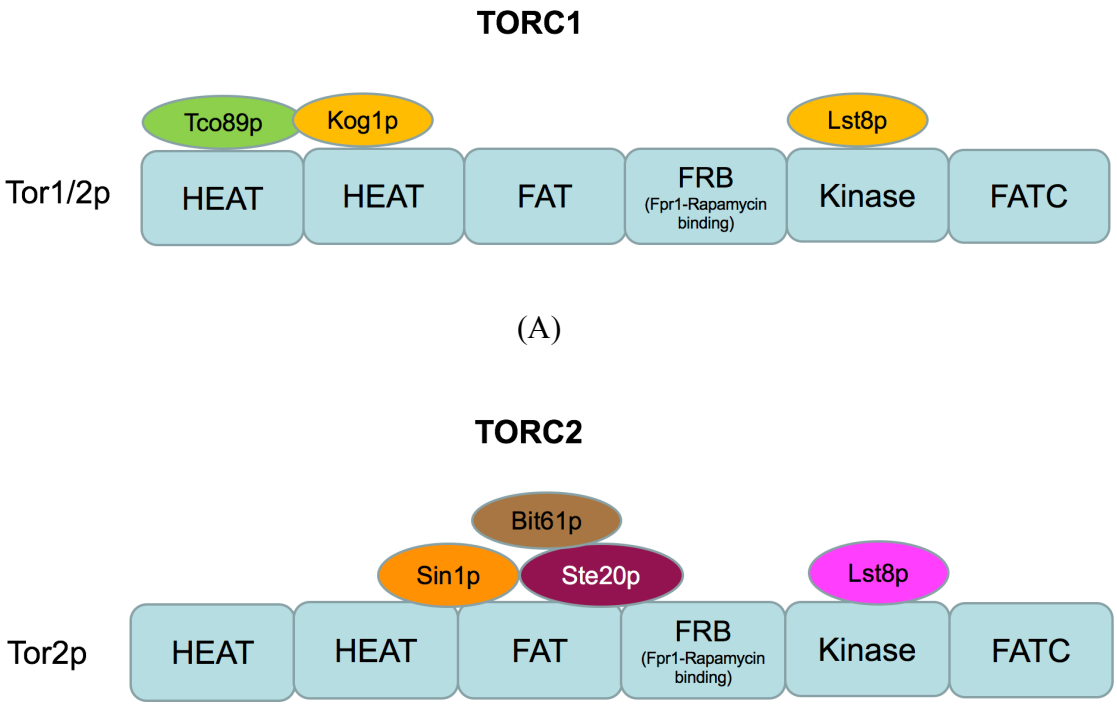


Figure 6.6 TOR complexes in *C. parasitica* modified from *S. cerevisiae* [113].

(A) represents the components of TORC1 and (B) represents the components of TORC2, modified from the TORC complexes in yeast [113].

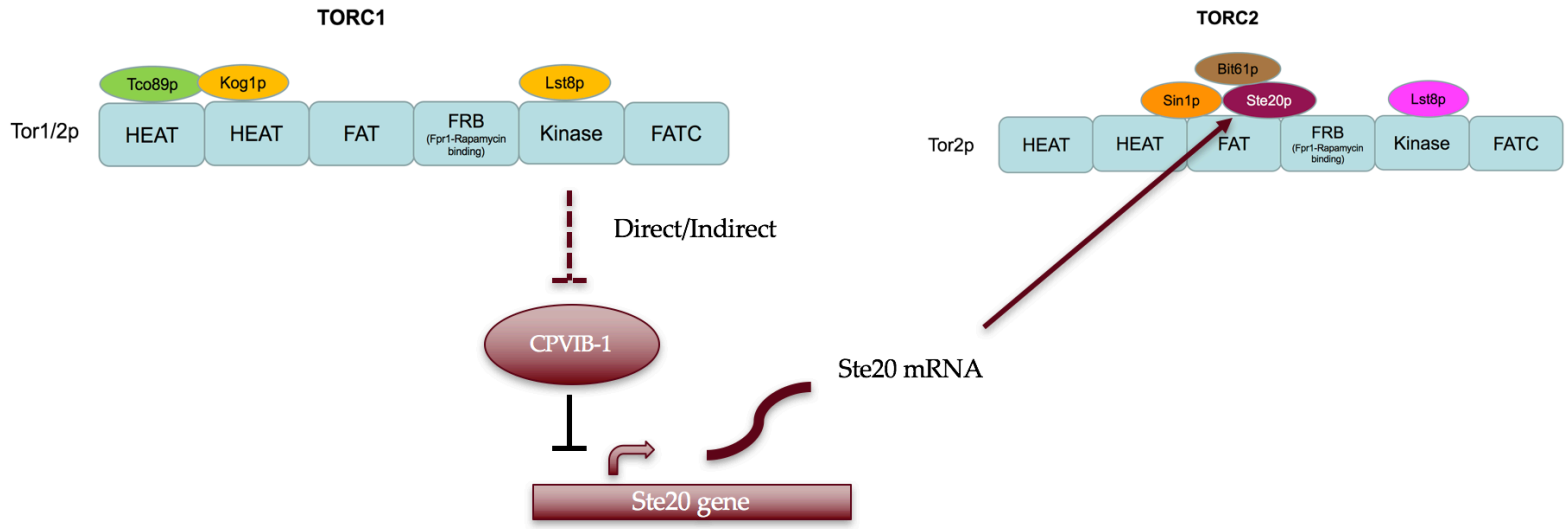


Figure 6.7 A revised model of TOR signaling that places CPVIB-1 between the TORC1 and TORC2 based on the results of this study.

REFERENCES

1. WA., M., *A new chestnut disease.* Torreya, 1906. **6**(9): p. 186-9.

2. Merkel, H.W., *A deadly fungus on the American chestnut.* NY Zool. Soc. Annu. Rep., 1906. **10**: p. 97-103.

3. Anderson, P.J. and W.H. Rankin, *Endothia canker of chestnut*. Vol. 347. 1914: Cornell University.

4. Gryzenhout, M., B.D. Wingfield, and M.J. Wingfield, *New taxonomic concepts for the important forest pathogen Cryphonectria parasitica and related fungi.* FEMS Microbiol Lett, 2006. **258**(2): p. 161-72.

5. Anagnostakis, S.L., *Chestnut blight: the classical problem of an introduced pathogen.* Mycologia, 1987. **79**(1): p. 23-37.

6. Anagnostakis, S.L., *Chestnut bark tannin assays and growth of chestnut blight fungus on extracted tannin.* Journal of chemical ecology, 1992. **18**(8): p. 1365-1373.

7. Anagnostakis, S.L. and B. Hillman, *Evolution of the chestnut tree and its blight.* Arnoldia, 1992. **52**(2): p. 2-10.

8. Dalgleish, H., et al., *Consequences of Shifts in Abundance and Distribution of American Chestnut for Restoration of a Foundation Forest Tree.* Forests, 2015. **7**(12): p. 4.

9. Heiniger, U. and D. Rigling, *Biological control of chestnut blight in Europe.*, in *Annu.Rev.Phytopathol.* 1994. p. 581-599.

10. Grente, J. and S. Sauret, *L'hypovirulence exclusive est-elle controllee par des determinants cytoplasmiques?* C.R. Acad. Sci. Paris Ser.D., 1969. **268**: p. 3173-3176.

11. Eusebio-Cope, A., et al., *The chestnut blight fungus for studies on virus/host and virus/virus interactions: from a natural to a model host.* Virology, 2015. **477**: p. 164-175.

12. Shapira, R., et al., *The contribution of defective RNAs to the complexity of viral-encoded double-stranded RNA populations present in hypovirulent strains of the chestnut blight fungus Cryphonectria parasitica.* EMBO J., 1991. **10**(4): p. 741-746.

13. Choi, G.H., R. Shapira, and D.L. Nuss, *Cotranslational autoproteolysis involved in gene expression from a double-stranded RNA genetic element associated with hypovirulence of the chestnut blight fungus.* Proc.Natl.Acad.Sci.U.S.A, 1991. **88**(4): p. 1167-1171.

14. Craven, M.G., et al., *Papain-like protease p29 as a symptom determinant encoded by a hypovirulence-associated virus of the chestnut blight fungus.* J.Virol., 1993. **67**(11): p. 6513-6521.

15. Shapira, R. and D.L. Nuss, *Gene expression by a hypovirulence-associated virus of the chestnut blight fungus involves two papain-like protease activities. Essential residues and cleavage site requirements for p48 autoproteolysis.* J.Biol.Chem., 1991. **266**(29): p. 19419-19425.

16. Anagnostakis, S.L. and P.R. Day, *Hypovirulence conversion in Endothia parasitica.* Phytopathology, 1979. **69**: p. 1226-1229.

17. Nuss, D.L., *Biological control of chestnut blight: an example of virus-mediated attenuation of fungal pathogenesis.* Microbiol.Rev., 1992. **56**(4): p. 561-576.

18. Milgroom, M.G. and P. Cortesi, *Biological control of chestnut blight with hypovirulence: a critical analysis.* Annu Rev Phytopathol, 2004. **42**: p. 311-38.

19. Biella, S., et al., *Programmed cell death correlates with virus transmission in a filamentous fungus.* Proceedings of the Royal Society of London *B*, 2002. **269**(1506): p. 2269-2276.

20. Zhang, D.X., et al., *Vegetative incompatibility loci with dedicated roles in allorecognition restrict mycovirus transmission in chestnut blight fungus.* Genetics, 2014. **197**(2): p. 701-14.

21. Paoletti, M. and S.J. Saupe, *Fungal incompatibility: evolutionary origin in pathogen defense?* Bioessays, 2009. **31**(11): p. 1201-10.

22. Saupe, S.J. and N.L. Glass, *Allelic specificity at the het-c heterokaryon incompatibility locus of Neurospora crassa is determined by a highly variable domain.* Genetics, 1997. **146**(4): p. 1299-309.

23. Cortesi, P., et al., *Genetic control of horizontal virus transmission in the chestnut blight fungus, Cryphonectria parasitica.* Genetics, 2001. **159**(1): p. 107-18.

24. Choi, G.H., et al., *Molecular characterization of vegetative incompatibility genes that restrict hypovirus transmission in the chestnut blight fungus Cryphonectria parasitica.* Genetics, 2012. **190**(1): p. 113-27.

25. Espagne, E., et al., *HET-E and HET-D belong to a new subfamily of WD40 proteins involved in vegetative incompatibility specificity in the fungus Podospora anserina.* Genetics, 2002. **161**(1): p. 71-81.

26. Kubisiak, T.L. and M.G. Milgroom, *Markers linked to vegetative incompatibility (vic) genes and a region of high heterogeneity and reduced recombination near the mating type locus (MAT) in Cryphonectria parasitica.* Fungal Genet Biol, 2006. **43**(6): p. 453-63.

27. Pinto, I., D.E. Ware, and M. Hampsey, *The yeast SUA7 gene encodes a homolog of human transcription factor TFIIB and is required for normal start site selection in vivo.* Cell, 1992. **68**(5): p. 977-88.

28. Galagan, J.E., et al., *The genome sequence of the filamentous fungus Neurospora crassa.* Nature, 2003. **422**(6934): p. 859-68.

29. Staben C, Y.C., *Neurospora crassa a mating-type region.* . Proceedings of the National Academy of Sciences, 1990. **87**(13): p. 4917-21.

30. Hutchison, E.A. and N.L. Glass, *Interplay Between NDT80 Homologs and the Protein Kinase IME-2 Regulates Sexual Development, but Not Meiosis, in Neurospora crassa.* Genetics, 2010.

31. Katz, M.E. and S. Cooper, *Extreme Diversity in the Regulation of Ndt80-Like Transcription Factors in Fungi.* G3 (Bethesda), 2015. **5**(12): p. 2783-92.

32. Xiang, Q. and N.L. Glass, *Identification of vib-1, a locus involved in vegetative incompatibility mediated by het-c in Neurospora crassa.* Genetics, 2002. **162**(1): p. 89-101.

33. Xiang, Q. and N.L. Glass, *The control of mating type heterokaryon incompatibility by vib-1, a locus involved in het-c heterokaryon incompatibility in Neurospora crassa.* Fungal Genet Biol, 2004. **41**(12): p. 1063-76.

34. Xiong, Y., J. Sun, and N.L. Glass, *VIB1, a link between glucose signaling and carbon catabolite repression, is essential for plant cell wall degradation by Neurospora crassa.* PLoS Genet, 2014. **10**(8): p. e1004500.

35. Rong, M., *Investigation of factors associated with vegetative incompatibility and virus transmission in cryphonectria parasitica*, in *Department of Biology*. 2011, New Mexico State University.

36. Nowrousian, M., *Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems.* Eukaryotic cell, 2010. **9**(9): p. 1300-1310.

37. Eddy, S.R., *What is a hidden Markov model?* Nature Biotechnology, 2004. **22**: p. 1315-1316.

38. Stanke, M., et al., *AUGUSTUS: a web server for gene finding in eukaryotes.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W309-12.

39. Yandell, M. and D. Ence, *A beginner's guide to eukaryotic genome annotation.* Nat Rev Genet, 2012. **13**(5): p. 329-42.

40. Feldmesser, E., et al., *Improving transcriptome construction in non-model organisms: integrating manual and automated gene definition in Emiliania huxleyi.* BMC genomics, 2014. **15**(1): p. 148.

41. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2008. **36**(Database issue): p. D25-30.

42. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

43. Wang, J., et al., *Identification of Novel Transcribed Regions in Zebrafish (Danio rerio) Using RNA-Sequencing.* PLoS One, 2016. **11**(7): p. e0160197.

44. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature protocols, 2012. **7**(3): p. 562.

45. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.* BMC bioinformatics, 2011. **12**(1): p. 491.

46. Korf, I., *SNAP: Semi-HMM-based Nucleic Acid Parser.* Ian Korf homepage: http://homepage/. mac. com/iankorf, 2013.

47. Lomsadze, A., et al., *Gene identification in novel eukaryotic genomes by self-training algorithm.* Nucleic acids research, 2005. **33**(20): p. 6494-6506.

48. Yandell, C.H.a.M., *MAKER2: an annotation pipeline and genome database management tool for second generation genome projects.* BMC Bioinformatics, 2011. **12**: p. 491.

49. Thrasher, A., et al., *Scaling up genome annotation using MAKER and work queue.* International Journal of Bioinformatics Research and Applications 2, 2014. **10**(4-5): p. 447-460.

50. Wang, Z., Y. Chen, and Y. Li, *A brief review of computational gene prediction methods.* Genomics, proteomics & bioinformatics, 2004. **2**(4): p. 216-221.

51. Allen, T.D., A.L. Dawe, and D.L. Nuss, *Use of cDNA microarrays to monitor transcriptional responses of the chestnut blight fungus Cryphonetria parasitica to infection by virulence-attenuating hypoviruses.* Eukaryotic Cell, 2003. **2**(6): p. 1253-1265.

52. Ghosh, S. and C.-K.K. Chan, *Analysis of RNA-Seq data using TopHat and Cufflinks*, in *Plant Bioinformatics*. 2016, Springer. p. 339-361.

53. Consortium, U., *UniProt: a hub for protein information.* Nucleic acids research, 2014. **43**(D1): p. D204-D212.

54. Horton, P., et al., *WoLF PSORT: protein localization predictor.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W585-7.

55. Knowledgebase, U., *a hub of integrated protein data Magrane Michele; Consortium Uniprot Database: the journal of biological databases and curation (2011), 2011 (), bar009 ISSN.* The UniProt Knowledgebase (UniProtKB) acts as a central hub of protein knowledge by providing a unified view of protein sequence and functional information. Manual and automatic annotation procedures are used to add data directly to the database while extensive cross-referencing to more than. **120**.

56. Lukashin, A.V. and M. Borodovsky, *GeneMark. hmm: new solutions for gene finding.* Nucleic acids research, 1998. **26**(4): p. 1107-1115.

57. Stanke, M. and B. Morgenstern, *AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints.* Nucleic acids research, 2005. **33**(suppl_2): p. W465-W467.

58. Stanke, M. and S. Waack, *Gene prediction with a hidden Markov model and a new intron submodel.* Bioinformatics, 2003. **19**(suppl_2): p. ii215-ii225.

59. Campbell, M.S., et al., *Genome Annotation and Curation Using MAKER and MAKER-P.* Curr Protoc Bioinformatics, 2014. **48**: p. 4 11 1-39.

60. Apweiler, R., et al., *The InterPro database, an integrated documentation resource for protein families, domains and functional sites.* Nucleic acids research, 2001. **29**(1): p. 37-40.

61. Apweiler, R., et al., *InterPro—an integrated documentation resource for protein families, domains and functional sites.* Bioinformatics, 2000. **16**(12): p. 1145-1150.

62. Quevillon, E., et al., *InterProScan: protein domains identifier.* Nucleic acids research, 2005. **33**(suppl_2): p. W116-W120.

63. Bell, J.A., et al., *Physical and genetic map of the mitochondrial genome of Cryphonectria parasitica Ep155.* Current genetics, 1996. **30**(1): p. 34-43.

64. 't Hoen, P.A., et al., *Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.* Nucleic acids research, 2008. **36**(21): p. e141-e141.

65. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome biology, 2016. **17**(1): p. 13.

66. Syme, R.A., et al., *Comprehensive Annotation of the Parastagonospora nodorum Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics.* PLoS One, 2016. **11**(2): p. e0147221.

67. Vleeshouwers, V.G. and R.P. Oliver, *Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens.* Molecular plant-microbe interactions, 2014. **27**(3): p. 196-206.

68. Liu, Y. and B. Schmidt, *Long read alignment based on maximal exact match seeds.* Bioinformatics, 2012. **28**(18): p. i318-i324.

69. Stothard, P. and D.S. Wishart, *Automated bacterial genome analysis and annotation.* Curr Opin Microbiol, 2006. **9**(5): p. 505-10.

70. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.

71. Markowitz, V.M., et al., *The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions.* Nucleic Acids Res, 2008. **36**(Database issue): p. D528-33.

72. Liu, X., X. Jian, and E. Boerwinkle, *dbNSFP v2. 0: a database of human non-synonymous SNVs and their functional predictions and annotations.* Human mutation, 2013. **34**(9).

73. Karen Eilbeck*, S.E.L., Christopher J Mungall†, Mark Yandell†, and R.D.a.M.A. Lincoln Stein‡, *The Sequence Ontology: a tool for the unification of genome annotations.* Genome Biology, 2005. **6**: p. R44.

74. Kalkatawi, M., I. Alam, and V.B. Bajic, *BEACON: automated tool for Bacterial GEnome Annotation ComparisON.* BMC Genomics, 2015. **16**: p. 616.

75. Liu Z, M.H., Goryanin I., *A semi-automated genome annotation comparison and integration scheme.* BMC Bioinformatics, 2013. **14**(1): p. 172.

76. Allemann, C., et al., *Genetic variation of Cryphonectria hypoviruses (CHV1) in Europe, assessed using restriction fragment length polymorphism (RFLP) markers.* Mol.Ecol., 1999. **8**(5): p. 843-854.

77. Cortesi, P. and M.G. Milgroom, *Genetics of vegetative incompatibility in cryphonectria parasitica.* Appl Environ Microbiol, 1998. **64**(8): p. 2988-94.

78. Colot, H.V., et al., *A high-throughput gene knockout procedure for Neurospora reveals functions for multiple transcription factors.* Proc. Natl. Acad. Sci. USA, 2006. **103**(27): p. 10352-10357.

79. Dementhon, K., G. Iyer, and N.L. Glass, *VIB-1 is required for expression of genes necessary for programmed cell death in Neurospora crassa.* Eukaryot Cell, 2006. **5**(12): p. 2161-73.

80. Nowrousian, M., *Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems.* Eukaryot Cell, 2010. **9**(9): p. 1300-10.

81. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

82. Andrews, S., *FastQC: a quality control tool for high throughput sequence data.* 2010.

83. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

84. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

85. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

86. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

87. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

88. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

89. Luo, W., et al., *GAGE: generally applicable gene set enrichment for pathway analysis.* BMC Bioinformatics, 2009. **10**: p. 161.

90. Luo, W. and C. Brouwer, *Pathview: an R/Bioconductor package for pathway-based data integration and visualization.* Bioinformatics, 2013. **29**(14): p. 1830-1.

91. JH, Z., *Pedigree-drawing with R and graphviz.* Bioinformatics, 2006. **22**(8): p. 1013-4.

92. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PLoS One, 2011. **6**(7): p. e21800.

93. Williams, C.R., et al., *Trimming of sequence reads alters RNA-Seq gene expression estimates.* BMC Bioinformatics, 2016. **17**: p. 103.

94. Lushchak, V.I., *Adaptive response to oxidative stress: Bacteria, fungi, plants and animals.* Comp Biochem Physiol C Toxicol Pharmacol, 2011. **153**(2): p. 175-90.

95. Smith, M.L., C.C. Gibbs, and M.G. Milgroom, *Heterokaryon incompatibility function of barrage-associated vegetative incompatibility genes (vic) in Cryphonectria parasitica.* Mycologia, 2006. **98**(1): p. 43-50.

96. Lomaev, D., et al., *The GAGA factor regulatory network: Identification of GAGA factor associated proteins.* PLoS One, 2017. **12**(3): p. e0173602.

97. Zhang, H., et al., *MgCRZ1, a transcription factor of Magnaporthe grisea, controls growth, development and is involved in full virulence.* FEMS Microbiol Lett, 2009. **293**(2): p. 160-9.

98. Schmidt, D., et al., *ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions.* Methods, 2009. **48**(3): p. 240-8.

99. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

100. Mundade, R., et al., *Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond.* Cell Cycle, 2014. **13**(18): p. 2847-52.

101. Desai, B., B. Myers, and S. Schreiber, *FKBP12-rapamycin-associated protein associates with mitochondria and senses osmotic stress via mitochondrial dysfunction.* Proc Natl Acad Sci, 2002. **99**(7): p. 4319-24.

102. Saxton, R.A. and D.M. Sabatini, *mTOR Signaling in Growth, Metabolism, and Disease.* Cell, 2017. **168**(6): p. 960-976.

103. Loewith, R. and M.N. Hall, *Target of rapamycin (TOR) in nutrient signaling and growth control.* Genetics, 2011. **189**(4): p. 1177-201.

104. Rohde, J., J. Heitman, and M.E. Cardenas, *The TOR kinases link nutrient sensing to cell growth.* J Biol Chem, 2001. **276**(13): p. 9583-6.

105. Dementhon, K., et al., *Rapamycin mimics the incompatibility reaction in the fungus Podospora anserina.* Eukaryot Cell, 2003. **2**(2): p. 238-46.

106. Schafer K, B.T., *Monoclonal anti-FLAG antibodies react with a new isoform of rat Mg2+ dependent protein phosphatase β.* . Biochemical and biophysical research communications., 1995. **207**(2): p. 708-14.

107. Bhowmick R, H.U., Chattopadhyay S, Nayak MK, Chawla-Sarkar M., *Rotavirus-encoded nonstructural protein 1 modulates cellular apoptotic machinery by targeting tumor suppressor protein p53.* Journal of virology, 2013. **87**(12): p. 6840-50.

108. Chambers AE, S.P., Randeva H, Banerjee S. , *Microvesicle-mediated release of soluble LH/hCG receptor (LHCGR) from transfected cells and placenta explants.* . Reproductive Biology and Endocrinology., 2011. **9**(1): p. 64.

109. de Castro, P.A., et al., *ChIP-seq reveals a role for CrzA in the Aspergillus fumigatus high-osmolarity glycerol response (HOG) signalling pathway.* Mol Microbiol, 2014. **94**(3): p. 655-74.

110. *<gaga factor.pdf>*.

111. Bejarano, F. and A. Busturia, *Function of the Trithorax-like gene during Drosophila development.* Dev Biol, 2004. **268**(2): p. 327-41.

112. Hecker, A., et al., *The Arabidopsis GAGA-Binding Factor BASIC PENTACYSTEINE6 Recruits the POLYCOMB-REPRESSIVE COMPLEX1 Component LIKE HETEROCHROMATIN PROTEIN1 to GAGA DNA Motifs.* Plant Physiol, 2015. **168**(3): p. 1013-24.

113. Inoue, Y. and W. Nomura, *TOR Signaling in Budding Yeast*, in *The Yeast Role in Medical Applications*. 2018, InTech.

114. Churchill, A.C.L., et al., *Transformation of the fungal pathogen Cryphonectria parasitica with a variety of heterologous plasmids.* Curr Genet, 1990. **17**: p. 25-31.

115. McLean, T.C., P.A. Hoskisson, and R.F. Seipke, *Coordinate Regulation of Antimycin and Candicidin Biosynthesis.* mSphere, 2016. **1**(6).

116. Anagnostakis, S.L., et al., *Hypovirus transmission to ascospore progeny by field-released transgenic hypovirulent strains of Cryphonectria parasitica.* Phytopathology, 1998. **88**(7): p. 598-604.

117. Patel, N., et al., *Use of the tetrazolium salt MTT to measure cell viability effects of the bacterial antagonist Lysobacter enzymogenes on the filamentous fungus Cryphonectria parasitica.* Antonie Van Leeuwenhoek, 2013. **103**(6): p. 1271-80.

118. Feng, J., et al., *Identifying ChIP-seq enrichment using MACS.* Nat Protoc, 2012. **7**(9): p. 1728-40.

119. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.

120. Muratani, M. and W.P. Tansey, *How the ubiquitin-proteasome system controls transcription.* Nat Rev Mol Cell Biol, 2003. **4**(3): p. 192-201.

121. Loh, Y.E., et al., *Bioinformatic Analysis for Profiling Drug-induced Chromatin Modification Landscapes in Mouse Brain Using ChIP-seq Data.* Bio Protoc, 2017. **7**(3).

122. Steinhauser, S., et al., *A comprehensive comparison of tools for differential ChIP-seq analysis.* Brief Bioinform, 2016. **17**(6): p. 953-966.

123. Lin, R. and H. Wang, *Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of Arabidopsis development.* Plant Physiol, 2004. **136**(4): p. 4010-22.

124. Kooiker, M., et al., *BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic Arabidopsis gene SEEDSTICK.* Plant Cell, 2005. **17**(3): p. 722-9.

125. King, G.J., et al., *The Arabidopsis B3 domain protein VERNALIZATION1 (VRN1) is involved in processes essential for development, with structural and mutational studies revealing its DNA-binding surface.* J Biol Chem, 2013. **288**(5): p. 3198-207.

126. Chang, V.H., et al., *Kruppel-like factor 10 expression as a prognostic indicator for pancreatic adenocarcinoma.* Am J Pathol, 2012. **181**(2): p. 423-30.

127. Erdman S, L.L., Malczynski M, Snyder M., *Pheromone-regulated genes required for yeast mating differentiation.* . The Journal of cell biology, 1998. **140**(3): p. 461-83.

128. Van Dyke, M.W., et al., *Stm1p, a G4 quadruplex and purine motif triplex nucleic acid-binding protein, interacts with ribosomes and subtelomeric Y' DNA in Saccharomyces cerevisiae.* J Biol Chem, 2004. **279**(23): p. 24323-33.

129. Ee, G. and N. Lehming, *How the ubiquitin proteasome system regulates the regulators of transcription.* Transcription, 2012. **3**(5): p. 235-9.

130. Heinemeyer, W., P.C. Ramos, and R.J. Dohmen, *The ultimate nanoscale mincer: assembly, structure and active sites of the 20S proteasome core.* Cell Mol Life Sci, 2004. **61**(13): p. 1562-78.

131. Bailey, T., et al., *Practical guidelines for the comprehensive analysis of ChIP-seq data.* PLoS Comput Biol, 2013. **9**(11): p. e1003326.

132. Biggin MD, T.R., *Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts.* . Cell, 1988. **53**(5): p. 699-711.

133. Durmowicz, M.C. and R.J. Maier, *Roles of HoxX and HoxA in biosynthesis of hydrogenase in Bradyrhizobium japonicum.* J Bacteriol, 1997. **179**(11): p. 3676-82.

134. Kim, S., et al., *Homeobox transcription factors are required for conidiation and appressorium development in the rice blast fungus Magnaporthe oryzae.* PLoS Genet, 2009. **5**(12): p. e1000757.

135. Lekva, T., et al., *Epithelial splicing regulator protein 1 and alternative splicing in somatotroph adenomas.* Endocrinology, 2013. **154**(9): p. 3331-43.

136. Li, Y., et al., *MoRic8 Is a novel component of G-protein signaling during plant infection by the rice blast fungus Magnaporthe oryzae.* Mol Plant Microbe Interact, 2010. **23**(3): p. 317-31.

137. Nagata Y, O.M., Nakata H, Shozaki Y, Kozasa T, Todokoro K., *A novel regulator of G-protein signaling bearing GAP activity for Gαi and Gαq in megakaryocytes.* Blood, 2001. **97**(10): p. 3051-60.

138. Kim, S., et al., *Combining ChIP-chip and expression profiling to model the MoCRZ1 mediated circuit for Ca/calcineurin signaling in the rice blast fungus.* PLoS Pathog, 2010. **6**(5): p. e1000909.

139. Ding, S.L., et al., *The tig1 histone deacetylase complex regulates infectious growth in the rice blast fungus Magnaporthe oryzae.* Plant Cell, 2010. **22**(7): p. 2495-508.

140. Lahaye A, S.H., Thines-Sempoux D, Foury F. , *PIF1: a DNA helicase in yeast mitochondria.* The EMBO Journal, 1991. **10**(4): p. 997-1007.

141. Rehmeyer, C.J., et al., *The telomere-linked helicase (TLH) gene family in Magnaporthe oryzae: revised gene structure reveals a novel TLH-specific protein motif.* Curr Genet, 2009. **55**(3): p. 253-62.

142. Sbia, M., et al., *Regulation of the yeast Ace2 transcription factor during the cell cycle.* J Biol Chem, 2008. **283**(17): p. 11135-45.

143. Marshall, K.R., et al., *The human apoptosis-inducing protein AMID is an oxidoreductase with a modified flavin cofactor and DNA binding activity.* J Biol Chem, 2005. **280**(35): p. 30735-40.

144. Wu, D., et al., *ChLae1 and ChVel1 regulate T-toxin production, virulence, oxidative stress response, and development of the maize pathogen Cochliobolus heterostrophus.* PLoS Pathog, 2012. **8**(2): p. e1002542.

145. Ramanujam, R. and N.I. Naqvi, *PdeH, a high-affinity cAMP phosphodiesterase, is a key regulator of asexual and pathogenic differentiation in Magnaporthe oryzae.* PLoS Pathog, 2010. **6**(5): p. e1000897.

146. Monod, M. and Z.M. Borg-von, *Secreted aspartic proteases as virulence factors of Candida species.* Biol Chem, 2002. **383**(7-8): p. 1087-93.

147. Hayashi, T., et al., *Rapamycin sensitivity of the Schizosaccharomyces pombe tor2 mutant and organization of two highly phosphorylated TOR complexes by specific and common subunits.* Genes Cells, 2007. **12**(12): p. 1357-70.

148. Ouzounis CA, K.P., *The past, present and future of genome-wide re-annotation.* Genome Biology, 2002. **3**(2): p. comment200-1.

149. Dalman, K., et al., *A genome-wide association study identifies genomic regions for virulence in the non-model organism Heterobasidion annosum s.s.* PLoS One, 2013. **8**(1): p. e53525.

150. Ren, Z., et al., *Improvements to the rice genome annotation through large-scale analysis of RNA-Seq and proteomics datasets.* . bioRxiv, 2018: p. 300426.

151. Bakke, P., et al., *Evaluation of three automated genome annotations for Halorhabdus utahensis.* PLoS One, 2009. **4**(7): p. e6291.

152. Wang, P., et al., *p53 domains: structure, oligomerization, and transformation.* Mol Cell Biol, 1994. **14**(8): p. 5182-91.

153. Katz ME, B.K., Yi G, Cooper S, Nonhebel HM, Gondro C., *A p53-like transcription factor similar to Ndt80 controls the response to nutrient stress in the filamentous fungus, Aspergillus nidulans.* F1000Research, 2013. **2**.

154. Chu S, D.J., Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I., *The transcriptional program of sporulation in budding yeast.* Science, 1998. **282**(5389): p. 699-705.

155. MacRae WD, B.F., Sibley S, Garven S, Gwynne DI, Arst HN, Davies RW., *Characterization of an Aspergillus nidulans genomic DNA fragment conferring phosphate-non-repressible acid-phosphatase activity.* Gene, 1993. **130**(2): p. 247-51.

156. Montano SP, C.M., Fingerman I, Pierce M, Vershon AK, Georgiadis MM., *Crystal structure of the DNA-binding domain from Ndt80, a transcriptional activator required for meiosis in yeast.* . Proceedings of the National Academy of Sciences, 2002. **99**(22): p. 14041-6.

157. Filtz, T.M., W.K. Vogel, and M. Leid, *Regulation of transcription factor activity by interconnected post-translational modifications.* Trends Pharmacol Sci, 2014. **35**(2): p. 76-85.

158. Bannister AJ, M.E., *Regulation of gene expression by transcription factor acetylation.* Cellular and Molecular Life Sciences CMLS, 2000. **57**(8-9): p. 1184-92.

159. Conaway RC, B.C., Conaway JW., *Emerging roles of ubiquitin in transcription regulation.* Science, 2002. **296**(5571): p. 1254-8.

160. Whitmarsh AJ, D.R., *Regulation of transcription factor function by phosphorylation.* Cellular and Molecular Life Sciences CMLS, 2000. **57**(8-9): p. 1172-83.

161. Jadhav, T. and M.W. Wooten, *Defining an Embedded Code for Protein Ubiquitination.* J Proteomics Bioinform, 2009. **2**: p. 316.

162. Chen, C.G., et al., *CaNdt80 is involved in drug resistance in Candida albicans by regulating CDR1.* Antimicrob Agents Chemother, 2004. **48**(12): p. 4505-12.

163. Nobile, C.J., et al., *A recently evolved transcriptional network controls biofilm development in Candida albicans.* Cell, 2012. **148**(1-2): p. 126-38.

164. Sellam, A., et al., *Role of transcription factor CaNdt80p in cell separation, hyphal growth, and virulence in Candida albicans.* Eukaryot Cell, 2010. **9**(4): p. 634-44.

165. Katz, M.E., K.A. Gray, and B.F. Cheetham, *The Aspergillus nidulans xprG (phoG) gene encodes a putative transcriptional activator involved in the response to nutrient limitation.* Fungal Genet Biol, 2006. **43**(3): p. 190-9.

166. Srivastava, A., A.S. Kumar, and R.K. Mishra, *Vertebrate GAF/ThPOK: emerging functions in chromatin architecture and transcriptional regulation.* Cell Mol Life Sci, 2018. **75**(4): p. 623-633.

167. Alvarez, B. and S. Moreno, *Fission yeast Tor2 promotes cell growth and represses cell differentiation.* J Cell Sci, 2006. **119**(Pt 21): p. 4475-85.

168. Matsuo, T., et al., *Loss of the TOR kinase Tor2 mimics nitrogen starvation and activates the sexual development pathway in fission yeast.* Mol Cell Biol, 2007. **27**(8): p. 3154-64.

169. Nonaka H, T.K., Hirano H, Fujiwara T, Kohno H, Umikawa M, Mino A, Takai Y. , *A downstream target of RHO1 small GTP-binding protein is PKC1, a homolog of protein kinase C, which leads to activation of the MAP kinase cascade in Saccharomyces cerevisiae.* . The EMBO journal, 1995. **14**(23): p. 5931-8.

170. Schmitz HP, L.A., Heinisch JJ., *Regulation of yeast protein kinase C activity by interaction with the small GTPase Rho1p through its amino-terminal HR1 domain.* Molecular microbiology, 2002. **44**(3): p. 829-40.

171. Heitman, J., et al., *FK 506-binding protein proline rotamase is a target for the immunosuppressive agent FK 506 in Saccharomyces cerevisiae.* Proceedings of the National Academy of Sciences, 1991. **88**(5): p. 1948-1952.

172. Koltin, Y., et al., *Rapamycin sensitivity in Saccharomyces cerevisiae is mediated by a peptidyl-prolyl cis-trans isomerase related to human FK506-binding protein.* Molecular and Cellular Biology, 1991. **11**(3): p. 1718-1723.

173. Martel, R., J. Klicius, and S. Galet, *Inhibition of the immune response by rapamycin, a new antifungal antibiotic.* Canadian journal of physiology and pharmacology, 1977. **55**(1): p. 48-51.

174. Heitman, J., N.R. Movva, and M.N. Hall, *Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast.* Science, 1991. **253**(5022): p. 905-909.

175. Kunz, J., et al., *Target of rapamycin in yeast, TOR2, is an essential phosphatidylinositol kinase homolog required for G1 progression.* Cell, 1993. **73**(3): p. 585-596.

176. Lorenz, M.C. and J. Heitman, *TOR mutations confer rapamycin resistance by preventing interaction with FKBP12-rapamycin.* Journal of Biological Chemistry, 1995. **270**(46): p. 27531-27537.

177. Stan, R., et al., *Interaction between FKBP12-rapamycin and TOR involves a conserved serine residue.* Journal of Biological Chemistry, 1994. **269**(51): p. 32027-32030.

178. Zheng, X.-F., et al., *TOR kinase domains are required for two distinct functions, only one of which is inhibited by rapamycin.* Cell, 1995. **82**(1): p. 121-130.

179. Helliwell, S.B., et al., *TOR1 and TOR2 are structurally and functionally similar but not identical phosphatidylinositol kinase homologues in yeast.* Molecular biology of the cell, 1994. **5**(1): p. 105-118.

180. Loewith, R., et al., *Two TOR complexes, only one of which is rapamycin sensitive, have distinct roles in cell growth control.* Molecular cell, 2002. **10**(3): p. 457-468.

181. Reinke, A., et al., *TOR complex 1 includes a novel component, Tco89p (YPL180w), and cooperates with Ssd1p to maintain cellular integrity in Saccharomyces cerevisiae.* Journal of Biological Chemistry, 2004. **279**(15): p. 14752-14762.

182. Wedaman, K.P., et al., *Tor kinases are in distinct membrane-associated protein complexes in Saccharomyces cerevisiae.* Molecular biology of the cell, 2003. **14**(3): p. 1204-1220.

183. Cameron, A.J., et al., *mTORC2 targets AGC kinases through Sin1-dependent recruitment.* Biochemical Journal, 2011. **439**(2): p. 287-297.

184. Gatherar, I., et al., *Identification of a novel gene hbrB required for polarised growth in Aspergillus nidulans.* Fungal Genetics and Biology, 2004. **41**(4): p. 463-471.

185. Liao, H.-C. and M.-Y. Chen, *Target of rapamycin complex 2 signals to downstream effector yeast protein kinase 2 (Ypk2) through adheres-voraciously-to-target-of-rapamycin-2 protein 1 (Avo1) in Saccharomyces cerevisiae.* Journal of Biological Chemistry, 2012. **287**(9): p. 6089-6099.

186. Wullschleger, S., et al., *Molecular organization of target of rapamycin complex 2.* Journal of Biological Chemistry, 2005. **280**(35): p. 30697-30704.

187. Laplante, M. and D.M. Sabatini, *mTOR signaling in growth control and disease.* Cell, 2012. **149**(2): p. 274-293.

188. Sabatini, D.M., *Twenty-five years of mTOR: Uncovering the link from nutrients to growth.* Proceedings of the National Academy of Sciences, 2017: p. 201716173.

189. Menand, B., et al., *Expression and disruption of the Arabidopsis TOR (target of rapamycin) gene.* Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 6422-6427.

APPENDIX A

ADDITIONAL TABLES AND FIGURES

Additional Tables and Figures.

Table A.1 Primers used for validation of the prediction of 27 predicted genes (2009-version and 2017-version).

Category	Annotation	Gene ID	Forward Primer (5'-3')	Reverse Primer (5'-3')
SIMILAR	2009	258862	CTGGATTAGCGACTGCATAGAG	CGGAACGATCGTCACTGTT
	2017	Ep155_U_T00006215_1	CGAGCTGCTCATGCCTAAT	GAGGTGACCTTCTCGAATGATG
	2009	356517	/	/
	2017	Ep155_U_T00006216_1	GGTGGCAACGCAAGAAC	CAGTTTGTCGAAGGCTTTGT
	2009	255909	CAGCATTGTGAAGTCCGTCTA	GATTCGAGGTGATCCTGCTATATT
	2017	Ep155_U_T00004868_1	GGACCAGCTTACCACACTTATG	GATTCGAGGTGATCCTGCTATATT
	2009	231803	/	/
	2017	Ep155_U_T00004505_1	GCTCTACCAGCCTTAAGGAATC	GTGATCCGGCAGAAACCA
	2009	355955	/	/
	2017	Ep155_U_T00004864_1	CGGAAGAACAGAAGCAACAAG	GTACCACGCTCAACCAGTAA
	2009	263897	/	/
	2017	Ep155_U_T00008600_1	GAGATTCGGCAGATCAGTGG	AATACGATGCTTCTCGCTCTC
	2009	58876	/	/
	2017	Ep155_U_T00000186_1	CCATCGTCATCGCCTTCA	CAGCTACAGTCCTCGCA

Table A.1 (continued)

Different	2009	356706	/	/
	2017	Ep155_U_T00006769_1	ATGTCTGCCCGCGATTATT	GAGACCAGACTTTCCACCATAG
	2009	98319	ATGAGGTCTCCATCCATCCT	CTTACCCTCAAAGTCGGAGAAG
	2017	Ep155_U_T00006769_1	TGATCGTCTTCGACTCAAACTG	GCTCTCCATAGCCACTGTTC
	2009	231853	GGTCTTCACCTGAAGCTTGT	CAAGACGGCCCTTGGTATT
	2017	Ep155_U_T00007748_1	/	/
	2009	102253	GTCTTGAACTGGAAGGACGAG	GCCGAGATATGCGTCGAAA
	2017	Ep155_U_T00007119_1	GCAATGTGGCTCAGTTTCTTTC	CTTCGTTTGGGCGCATTTC
	2009	357202	/	/
	2017	Ep155_U_T00007829_1	GCAGAGGACTACGACTATGAAG	GCCTTATTCCCACCCATGA
	2009	261603	AAGGGCAAGAAGAGCAGAAG	GCATCGATCTACCAGCATGA
	2017	Ep155_U_T00007774_1	GATGCGGAGATGTGCATTG	GCATCGATCTACCAGCATGA
	2009	346358	GCTGTTTGTGGTCTCGTTATTC	GGTCGACCCTCACTTCAC
	2017	Ep155_U_T00005807_1	GGTCGACCCTCACTTCAC	CAGGAAAGACTCCGGTATGTAG
	2009	245160	/	/
	2017	Ep155_U_T00000872_1	ATGTCGCAGGGCTTGTC	GACAGATCCTGTTCCTCCTTC
	2009	357444	/	/
	2017	Ep155_U_T00008473_1	CTTCTTCTTCCCTCTCCATCAC	CATCTGGTCCTCGTCAAACA
	2009	222652	TCTTGATACAGGCGTTGATCTG	CTCTGGGCACAGTTGTGTAA
	2017	Ep155_U_T00007540_1	TGTTATCGCGAAGGGCTATC	GCGAGTATGAGCACCTCTTT
	2009	346810	TCTCTATTCCTGACTCGTCTCC	CCAGGCTTGTCTCGATCAAA
	2017	Ep155_U_T00006866_1	TCTCTATTCCTGACTCGTCTCC	CCGCTTGATCCTTCCTCTTTAT

Table A.1 (continued)

	2009	260426	GGTCGCGCAAGGTTAAATG	CTACAGATGAACGAGGTATGGG
	2017	Ep155_U_T00006108_1	TCACTGGTTTGGGTGTTAGG	CTACAGATGAACGAGGTATGGG
	2009	94097	AGGAGGTGGAGCGTCTT	AGCACATTCGCCTTGTCT
	2017	Ep155_U_T00007212_1	AGGAGGTGGAGCGTCTT	AGCACATTCGCCTTGTCT
	2009	322230	/	/
	2017	Ep155_U_T00005405_1	ATGAACACGACCTTACTACGAG	GTGCCAACCGATCATGTAGA
	2009	231988	/	/
	2017	Ep155_U_T00001901_1	GGACAGGCAGAAGACATCAA	GGGTTATTCAGAGCCCATCTAC
Noexist	2009	65946	TGCAGTCTAACCTATCAGCAATC	GAACTCCCTGCGGATACTTG
	2009	241925	GAGCGTCGTGGTGTCATT AT	GAA CCA TCA GAG CAG AGAAGAG
	2009	75444	TGATCCTGTTAGCCTTGCTATT	CCAATGGGTTGGTAGTGTATCT
	2009	334967	ACTATTTGCCTTGCAGGTCTT	GCAGTGCATTGGCCTCT
	2009	71384	GCCTTGTATCTTCTTACGGATCTT	ATGCTCATGCTTCTGTGCTAT
	2009	68001	ATGGTGGGCAGCCTTTG	GTCAATGAATATGGTGCAGAGGA

"/" represents that it is not necessary to design primer for this version to be able using diagnostic PCR to validate the accuracy of the gene from both versions.

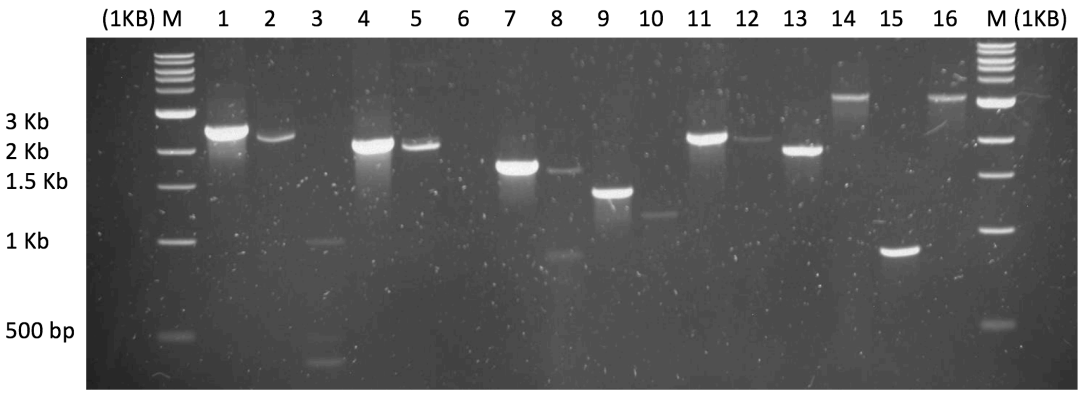


Figure A.1 Figure 1 of agarose gel pictures for validation of 27 predicted gene models (2009-version and 2017-version) using PCR.

(1KB) M is 1Kb size ladder, and lane 1, 4, 7, 9, 11 and 14 are the genomic DNA product of gene1, gene4, gene8 (2017-version), gene8 (2009-version), gene10 and gene17, separately. The other lanes' information is listed in the Table 3.3.

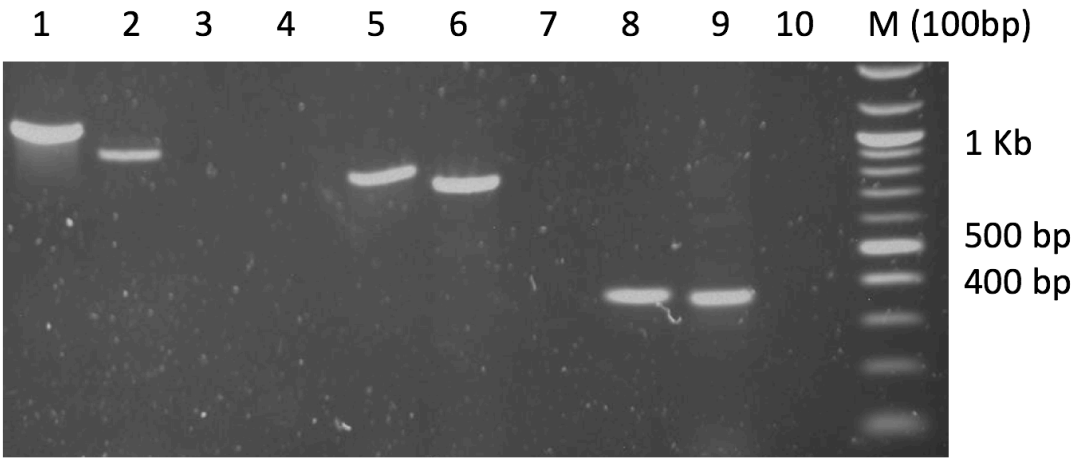


Figure A.2 Figure 2 of agarose gel pictures for validation of 27 predicted gene models (2009-version and 2017-version) using PCR.

lane 1, 3, 5 and 8 are the genomic DNA product of gene3(2009-version), gene3(2017-version), gene5 and gene7, separately, and M (100bp) is 100bp size ladder. The other lanes' information is listed in the Table 3.3.

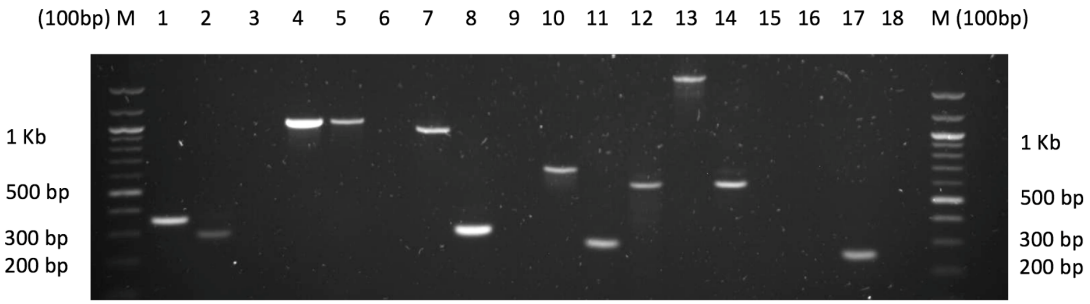


Figure A.3 Figure 3 of agarose gel pictures for validation of 27 predicted gene models (2009-version and 2017-version) using PCR.

(100 bp) M is 1Kb size ladder, and lane 1, 4, 7, 10, 13 and 16 are the genomic DNA product of gene2, gene9, gene11, gene12, gene13 and gene14, separately. The other lanes' information is listed in the Table 3.3.

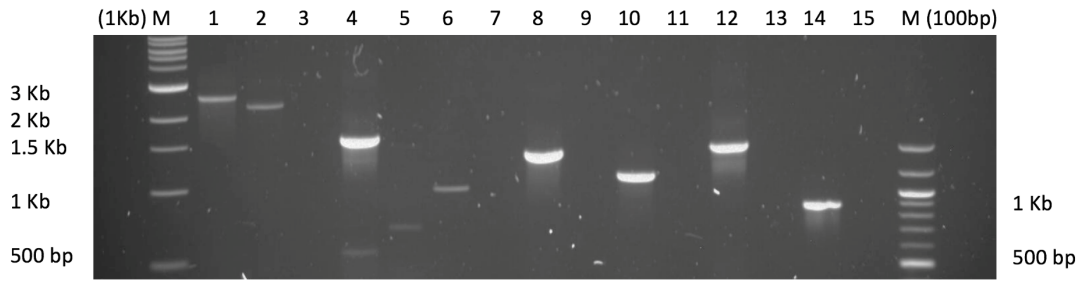


Figure A.4 Figure 4 of agarose gel pictures for validation of 27 predicted gene models (2009-version and 2017-version) using PCR.

(1KB) M is 1Kb size ladder, and lane 1, 4, 6, 8, 10, 12 and 14 are the genomic DNA product of gene 21, gene 23, gene24, gene25, gene26, gene27 and gene22, separately, and in the end M(100bp) is 100bp size ladder. The other lanes' information is listed in the Table 3.3.

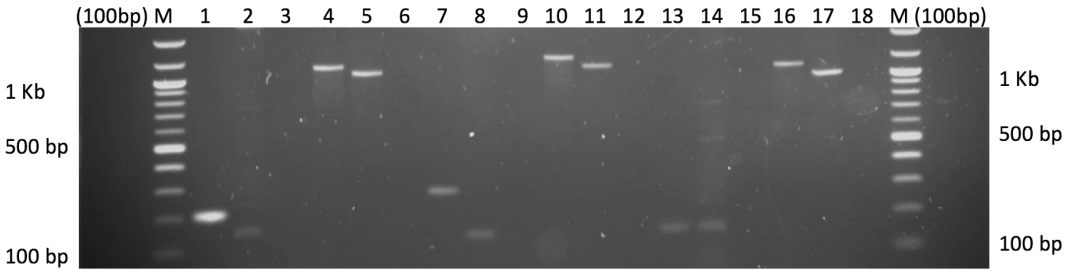


Figure A.5 Figure 5 of agarose gel pictures for validation of 27 predicted gene models (2009-version and 2017-version) using PCR.

(100bp) M is 100bp size ladder, and lane 1, 4, 7, 10, 13 and 16 are the genomic DNA product of gene6, gene15, gene16, gene18, gene19 and gene20, separately. The other lanes' information is listed in the Table 3.3.